A Theoretical Study of the Tryptophan Synthase Enzyme Reaction Network

---

Dissertation
zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer. nat.)

im Fach Chemie

Spezialisierung: Physikalische und theoretische Chemie

Eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität
zu Berlin

von
Dimitri Loutchko

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät
Prof. Dr. Elmar Kulke

---

1. Gutachter: Prof. Dr. Gerhard Ertl
2. Gutachter: Prof. Dr. Klaus Rademann
3. Gutachter: Prof. Dr. Yannick de Decker

Tag der mündlichen Prüfung: 09.07.2018

Abstract

The channeling enzyme tryptophan synthase provides a paradigmatic example of a chemical nanomachine. It catalyzes the biosynthesis of tryptophan from serine and indole glycerol phosphate. As a single macromolecule, it possesses two distinct catalytic subunits and implements 13 different elementary reaction steps. A complex pattern of allosteric regulation is involved in its operation. The catalytic activity in a subunit is enhanced or inhibited depending on the state of the other subunit. The gates controlling arrival and release of the ligands can become open or closed depending on the chemical states. The intermediate product indole is directly channeled within the protein from one subunit to another, so that it is never released into the solution around it.

In this thesis, the first single-molecule kinetic model of the enzyme is proposed and analyzed. All its transition rate constants are extracted from available experimental data, and thus, no fitting parameters are employed. Numerical simulations reveal strong correlations in the states of the active centers and the emergent synchronization of intramolecular processes in tryptophan synthase. Moreover, the effects of allosteric interactions are studied using modified *in silico* models with permanent and without any allosteric activations. The unmodified model of the native enzyme with transient activations significantly outperforms both modified models in terms of mean turnover times. An explanation is derived from the comparison of turnover time distributions showing a desynchronization of the two subunits in the modified models leading to cycles with long turnover times.

Thermodynamic data is used to calculate the rate constant for the reverse indole channeling, which has not been observed in experiments thus far. Using the fully reversible single-molecule model, the stochastic thermodynamics of the enzyme is examined. The Gibbs energy landscape of the internal molecular states is determined and the production of entropy and its flow within the enzyme are analyzed. The current methods describing information exchange in bipartite systems are extended to arbitrary Markov networks and applied to the kinetic model of tryptophan synthase. They allow the characterization of the information exchange between the subunits resulting from allosteric cross-regulations and channeling.

The last part of this work is focused on chemical reaction networks of metabolites and enzymes. Algebraic semigroup models are constructed based on a formalism that emphasizes the catalytic function of reactants within the network. These models admit a notion of successive and simultaneous functions not only of individual enzymes, but of any subnetwork. This includes the catalytic function of the whole reaction network on itself. The function is then used to decide whether the network is self-sustaining and a natural discrete dynamics is utilized to identify the maximal self-sustaining subnetwork. Then, a correspondence between coarse-graining procedures and semigroup congruences respecting the functional structure is established. A family of congruences that leads to a rather unusual coarse-graining is constructed: The network is covered with local patches in a way that the local information on the network is fully retained, but the environment of each patch is no longer resolved. Whereas classical coarse-graining procedures would fix a particular local patch and delete detailed information about the environment, the algebraic approach keeps the structure of all local patches and allows the interaction of functions within distinct patches.

Zusammenfassung

Das Enzym Tryptophan Synthase ist ein ausgezeichnetes Beispiel einer molekularen Fabrik auf der Nanoskala. Es katalysiert die Biosynthese der essentiellen Aminosäure Tryptophan aus Serin und Indol-glycerolphosphat. Der katalytische Zyklus des Moleküls beinhaltet mindestens 13 Elementarreaktionen, die in den katalytischen Zentren seiner zwei Untereinheiten stattfinden. Die Katalyse beruht zudem auf zahlreichen allosterischen Wechselwirkungen sowie der Übertragung des Intermediats Indol durch einen intramolekularen Tunnel.

In dieser Arbeit wird das erste kinetische Modell eines einzelnen Tryptophan Synthase Moleküls konstruiert und analysiert. Sämtliche Reaktionskonstanten sind aus der Literatur bekannt, wo-durch das Modell keine freien Parameter enthält. Numerische Simulationen zeigen starke Korrelationen zwischen den Zuständen der Katalysezentren sowie die Ausbildung von Synchronisation zwischen den intramolekularen Prozessen im Enzym. Des Weiteren werden die Effekte der allosterischen Wechselwirkungen durch den Einsatz von Modifikationen des Modells *in silico*, welche die Wechselwirkungen vollständig unterdrücken bzw. permanent aktivieren, untersucht. Es zeigt sich, dass das native Enzym eine erhelblich größere Reaktionsgeschwindigkeit aufweist als beide Modifikationen. Durch eine Analyse der Histogramme der Umsatzzeiten einzelner Zyklen lässt sich diese Beobachtung auf eine selten auftretende Desynchronisation der Katalysezyklen in den Untereinheiten, welche zu sehr langen Umsatzzeiten führt, zurückführen.

Die thermodynamischen Eigenschaften des Modells werden mithilfe der stochastischen Thermodynamik untersucht. Zunächst wird die experimentell unzugängliche Reaktionskonstante für die Rückübertragung des Indols aus thermodynamischen Messdaten rekonstuiert. Die freie Enthalphie aller chemischen Zustände des Moleküls, die Entropieproduktion sowie der Entropiefluss werden berechnet. Methoden, die den Informationsaustausch in bipartiten Markovnetzwerken charakterisieren, werden auf beliebige Markovnetzwerke verallgemeinert. Ihre Anwendung auf das kinetische Modell der Tryptophan Synthase führt zu einer Charakterisierung des Informationsaustauschs zwischen den Untereinheiten des Enzyms.

Der abschließende Teil der Arbeit befasst sich mit chemischen Reaktionsnetzwerken von Metaboliten und Enzymen. Ausgehend von einem Formalismus, der die katalytische Funktion von Reaktanten des Netzwerks hervorhebt, werden algebraische Modelle konstruiert. Es handelt sich dabei um Halbgruppen, welche aufeinanderfolgende und simultane katalytische Funktionen von Enzymen und von Unternetzwerken erfassen. Die Funktion des Netzwerkes auf sich selbst wird genutzt, um hinreichende und notwendige Bedingungen für seine Selbsterhaltung zu formulieren. Die Definition einer natürlichen Dynamik auf den Netzwerken erlaubt auch die Bestimmung des maximalen selbsterhaltenden Unternetzwerkes. Anschließend werden die algebraischen Modelle dazu genutzt, um eine Korrespondenz zwischen Halbgruppenkongruenzen und Skalenübergängen auf den Reaktionsnetzwerken herzustellen. Insbesondere wird eine Art von Kongruenzen erörtert, welche dem Ausspuren der globalen Struktur des Netzwerkes unter vollständiger Beibehaltung seiner lokalen Komponenten entspicht. Während klassische Techniken eine bestimmte lokale Komponente fixieren und sämtliche Informationen über ihre Umgebung ausspuren, sind bei dem algebraischen Verfahren alle lokalen Komponenten zugleich sichtbar und eine Verknüpfung von Funktionen aus verschiedenen Komponenten ist problemlos möglich.

Contents

List of Figures

List of Tables

Acknowledgement

First and foremost, I would like to express my deep gratitude to my thesis advisors Prof. Gerhard Ertl and Prof. Alexander S. Mikhailov.

Prof. Mikhailov has spent many hours of his valuable time on teaching me how to write scientific texts, prepare talks and posters and of course on interesting discussions. He is a true artist when it comes to breaking down complicated phenomena to their core and constructing elegant models that precisely capture the essence of the phenomenon without any unnecessary distractive features. Watching his process was an invaluable experience for me. Prof. Mikhailov pushed me when it was necessary, but also gave me the freedom to wander off and explore on my own for months. Throughout my time as a student in his group, he was always there to help, to discuss and to guide.

Prof. Ertl had a magical way of guidance. Each and every conversation with him gave me a huge motivational boost to proceed with the project we had discussed. He gave me an enormous freedom to pursue my ideas, but also set wise boundaries based on his experience. I am especially thankful to him for the consistent support of my studies in mathematics parallel to the thesis work. Through this support I could acquire the technical skills desperately needed for the work as a theorist. To me, Prof. Ertl is great example of a successful scientist and, at the same time, a great role model concerning human behavior. He takes great care of this staff and is as concerned about their well-being at least as much as about their scientific success.

I am thankful to Maximilian Görgner with whom I had an exciting time at the FHI learning stochastic thermodynamics together spiced up with discussions about science, philosophy and life in general. I have very much enjoyed the time spent in Brussels with Didier Gonze working on the model of tryptophan synthase. I am grateful for the invitation of Holger Flechsig to spend a month in Hiroshima filled with interesting discussions about science and many other things. I would also like to thank Prof. Hisao Hayakawa for his hospitality in Kyoto, where I had all the time and freedom to think about the possibility of algebraic models in biology. With gratitude I think about Amartya Sarkar, Holger Flechsig, Jeff Noel and all other members of the complex systems group who created a pleasant working atmosphere. I am much obliged to Prof. Klaus Rademann, who supported my thesis work at the FHI.

I am deeply thankful to my parents and my brother, who accepted that I spent much of my spare time and weekends to work on the thesis rather than joining them for social events. Yet, they have always been there for me when I needed help or advice. Not only for the last four years, but for my entire life my parents have always been giving and caring without ever asking for a return. I would like to dedicate chapter 5 of this work to them.

Introduction

Historically, the understanding of biological systems has been successively improved by the interplay of system reduction and integration of the reduced pieces. The reduction was, in most cases, enabled through the refinement of experimental techniques and the resulting possibility to observe smaller constituents of the system such as organs, cells, cell organelles, protein complexes, metabolites and ultimately the structure of DNA. Such constituents form a hierarchy resulting from the inclusion of smaller parts into larger structures. For example, cell organelles are included in cells and cells are included in organs. The goal of integration is concerned with the reconstruction of the properties of a particular structure from the properties of its lower level constituents. Each time new lower level structures have been discovered, the scientific community has spent much effort on formulating theories that achieve the integration of the newly found structures.

However, until the advent of molecular biology, for the lower level structures such as cells and cell organelles general interaction laws could not be formulated. Only phenomenological models adjusted to the respective experimental situation with experimentally determined parameters were available. Molecular biology, for the first time, allowed to envision that precise statements about biological systems could be made based on *first principles*. After all, the exact physical laws governing the structure and dynamics of molecules had been discovered in the early 20th century. The experimental accessibility - and thus the potential knowledge - of all molecular parts of an organism marks an important milestone in the biological science: it is the completion of the reduction program and the end of a conceptual dichotomy between reductive and integrative ways of thought.

Meanwhile, the integration of lower level structures is far from being completed and is a main driving force in the life sciences. It is a recurring theme in numerous publications, where complex behavior is explained in terms of interactions of simpler lower-level constituents. The integrative branch of molecular biology is now known as systems biology. It seeks to combine high-throughput data on the numbers, interactions and even time-evolution of metabolites, proteins, lipids, mRNA and DNA in a cell in order to develop detailed *in silico* models of the whole cell.

A remarkable success of systems biology is the identification of the molecular mechanisms controlling the circadian rhythm, awarded the Nobel Prize in Physiology and Medicine in 2017. In gene knockout experiments, Benzer and Konopka were able to identify a single gene (named period) whose knockout disrupted the circadian rhythm in fruit flies. Later, Hall, Rosbash and Young could show that the protein encoded by the gene (also called period) inhibits the transcription of its own gene and thereby forms a feedback loop. The period protein is degraded through the influence of sunlight and therefore

its concentration fluctuates in a 24-hour rhythm driven by the day and night cycle: The concentration increases during the night (up to some threshold value controlled by the feedback loop) and decreases during the day (again to some threshold when degradation rate and synthesis rate cancel each other). In meticulous experimental work, the Nobel laureates were able to identify other genes and proteins that stabilize the regulatory network and control the entry of period into the nucleus. One of the many fascinating aspects about this work is the successful integration of simple chemical reactions governed by standard rate equations of a set of chemicals into a reaction network that has a significant influence on all the levels of organization within the organism: The circadian rhythm affects chemical characteristics such as hormone levels and metabolism, physical characteristics such as body temperature and blood pressure and even medicinal characteristics such as the desire for sleep, coordination, reaction times and mood. This means that the processes influenced by the reaction network based on period span a large interval of time and length scales emerging from the small time and length scales of individual chemical reactions involved in the network. In this regard, it is interesting to note that the circadian rhythm within each organism is controlled by a physical process on an astronomical scale, namely the earth's rotation with respect to the sun.

The connection of processes on different time and length scales is becoming an increasingly important theme in the life sciences: While in the example discussed above, the connection between the period reaction network and higher-scale properties has not been made quantitative, there have been remarkable achievements in constructing quantitative *in silico* multiscale models. An outstanding example is a series of models of the human heart constructed by Noble *et al.* [1, 2]. Such models include functionally important genes, proteins, metabolites and many details on ion channels at the molecular level. These are included in models of all the main types of cardiac myocytes, which in turn are used in three-dimensional reconstructions of the whole organ as an elastic object paying attention to fiber orientation, sheet structure and the heart nervous system. Using such advanced models, many pathological states of the heart could be reproduced based on changes in the protein composition, drug interactions, or mutations of the ion channels. Moreover, it was possible to study the influence of the heart contraction on the electrical state of the heart, giving unexpected results on the connection to changes in cell volume. Along the same lines, arrhythmic behavior was successfully reproduced from models of the metabolic and electrophysiological processes following energy deprivation.

The period reaction network governing the circadian rhythm and the multiscale heart models each represent a major theme in system biological thought: At the molecular level, models of reaction networks of metabolites (called the metabolome), interaction networks of proteins (proteome), gene regulatory networks (genome) and mRNA expression levels are being integrated to determine mechanisms and regulatory motives within such networks. Such models are based on large amounts of quantitative and qualitative data using high-throughput techniques that simultaneously monitor the cellular concentrations of a large number of different chemical species. Modern techniques even allow time-resolved data to be obtained. However, such approaches are inherently weak at capturing the emergence of and interactions with larger structures within an organism. In the example of the heart model, membranes, cells and the three-dimensional structure of the heart were not deduced from the respective molecular interaction networks, but added "by

hand". Moreover, not the full reaction and interaction networks of molecules were taken into account, but only those important for the higher-scale processes under consideration. This approach to systems biology is more an "artful crafting" of suitable models and less a "black-box" approach based on a fixed set of rules. Indeed, many prominent scientists such as Sydney Brenner [3], Dennis Noble [4] and Laurent Nottale [5] hold the opinion that there is no preferred scale of causation in nature and that neither the genetic code nor the molecular interaction network of organisms therefore contain a sufficient description of the organism.

Such problems are already present at one of the "lower levels" of organization, within individual proteins and their interactions networks. Does the understanding of individual proteins provide deeper understanding of protein-protein interactions? How is the catalytic mechanism of a protein in diluted solutions *in vitro* or *in silico* related to its function *in vivo*? How important is the role of protein complexes when integrating high-throughput data without any *a priori* information on such complexes?

Proteins can be thought as the executive power of the cell. They carry out almost all functions in the living cell that involve manipulation and modification of the chemical and physical constitution of the cell or its environment. Enzymes catalyze most of the chemical reactions inside the cell. Through kinetic control they enable metabolic reactions to take place in a controlled manner at appropriate rates. Moreover, key steps such as transcription, splicing and translation are carried out by large complexes. Motor proteins transport cargo in the cytoplasm or through the cell membrane and perform the various mechanical motions such as bacterial flagellar locomotion or muscle contraction in higher organisms. Proteins are crucial for the control of cellular processes. In particular, they are involved in the responses to external stimuli through signaling networks: Receptor proteins at the cell surface detect stimuli (e.g. from nutrients, poisons or hormones, but also mechanical stress) and initiate a cascaded response. Therein, several messenger proteins from a network reminiscent of a calculatory circuit including feedback control and amplification mechanisms. The circuit either directly initiates a response or leads to changes in protein biosynthesis through appropriate transcription factors.

All the processes just described heavily rely on the interaction between proteins - either within complexes or networks. A well-known example of an enzyme complex is the ribosome, consisting of the small and large subunits, ribosomal RNA and a variety of additional ribosomal proteins. Even larger structures are focal adhesions with over 50 proteins [6] or the spliceosome including over 200 proteins [7]. These complexes are sufficiently stable and the components well enough known that they can be studied *in vitro* or can already been observed *in vivo* using classical optical methods. However, the exact composition of these complexes varies dynamically in the living cell. For example, the number of proteins making up the spliceosome is known to vary by up to 60 between different functional states [8]. Such observations on large and well-known complexes seem to be just the tip of the iceberg concerning the role of enzyme complexes within living organisms. There is a growing volume of evidence suggesting that many biochemical reactions within a cell are catalyzed by multi-enzyme complexes with poorly understood and highly dynamic higher order structure [9, 10, 11, 12, 13, 14, 15]. These complexes can implement entire metabolic pathways or significant parts of them. Within a complex,

intermediate products can be directly channeled [10, 11] to other enzymes for further processing, resembling the operation of an industrial conveyor belt. Moreover, different enzymes in a complex are usually coupled through allosteric regulatory loops [15]. Because of product channeling and multiple allosteric interactions, a complex can operate in a synchronous manner, exhibiting strong correlations in the turnover cycles of involved enzymes. Experimental investigations of multi-enzyme complexes encounter difficulties because the complexes are often transient and only exist *in vivo* [12].

An interesting class of enzymes are channeling enzymes [16, 17] (see also review [18]). They are similar in their properties to multi-enzyme complexes, but, in contrast, are smaller and have a well defined structure. A prototypical example of a channeling enzyme is tryptophan synthase [19] (introduced in detail in section 1.1). It catalyzes the biosynthesis of the essential amino acid tryptophan from serine and indole glycerol phosphate (IGP). This enzyme is employed by all bacteria, plants, fungi, but not by higher organisms and thus, can be a target for the development of antibiotics [20]. Its substrate IGP is scarce inside the cell and, therefore, high catalytic efficiency is required. Furthermore, an intermediate product (indole) of the synthesis reaction is hydrophobic and can easily escape through the cell membrane. Therefore, its release into the cytoplasm must be avoided. Nature has found an elegant solution for these constraints. The entire synthesis encompassing 13 elementary reaction steps is performed within the enzyme with two different catalytic centers and the intermediate indole is channeled within the protein from one center to another. Thus, tryptophan synthase is a model for larger and more difficult to access protein complexes.

In chapter 2, a *single-molecule* model of tryptophan synthase is constructed. It takes into account correlations between the states of the two catalytic centers arising through substrate channeling and mutual allosteric regulation. The stochastic model is formulated in terms of a Markov network. Because of the extensive experimental data available, all relevant microscopic rate constants in the model could be directly deduced from the data, so that no fitting parameters have been employed. Numerical simulations yield direct evidence of the presence of strong correlations and intramolecular synchronization of chemical processes in tryptophan synthase. They also allow to analyze the role of allosteric regulations in raising the catalytic efficiency of this enzyme. This work has been published in [21].

In chapter 3, the constructed Markov transition network is studied using the theory of stochastic thermodynamics for the operation of a single enzyme. Thereby, additional calorimetric data is used to determine the rate constant for reverse channeling that has not been experimentally observed. The energy landscape is constructed and an analysis of the entropy production and entropy flow within the enzyme in the nonequilibrium state corresponding to physiological conditions is performed.

Chapter 4 is focused on the information theoretic aspects of allosteric interactions between the two enzyme subunits and on the information effects of channeling events. Recently, a theory of information transfer in bipartite Markov networks has been constructed [22, 23, 24]. Bipartite Markov networks are networks whose state space can be factored as a product space $A \times B$ of two subsystems A and B such that all transitions

change either the state of the A-subsystem or of the B-subsystem, but not both at the same time. The Markov network models of allosteric proteins have exactly this structure: The A-subsystem is the catalytic site and the B-subsystem is the allosteric site. A catalyzed reaction changes only the A-state and the binding or unbinding of some allosteric effector changes only the B-state. The allosteric interaction entails an effect of the B-state on the catalytic rates of the A-subsystem. This effect is made quantitative in the theory of information thermodynamics and, thus, it is straightforward to apply the theory to allosteric proteins. However, when mass transfer between the subsystem A and B takes place, there is no longer a bipartite structure, because substances leaving one subsystem immediately arrive in the other subsystem. In such cases, application of the theory is not straightforward, but it can be extended. This is done in section 4.1. As an illustration, the extended theory is applied to tryptophan synthase, which has both allosteric interactions between its two subunits and mass transfer due to indole channeling. The work presented in chapters 3 and 4 has been published in [25].

Chapter 5 takes a more general perspective on chemical reaction networks. The reaction networks are modeled by finite and discrete state spaces as in the case of the tryptophan synthase model. However, the states correspond to sets of metabolites and not to individual states of a single enzyme. As described in the first paragraphs of this introduction, high-throughput techniques generate large amounts of data on particular levels of organization, in particular, on reaction networks of metabolites, interaction networks of proteins and genetic regulatory networks. The connection between this data and the hierarchical organization of biological systems across many scales is an omnipresent theme in modern systems biology, which has fascinated this author ever since he became aware of it. The methods in chapter 5 are a non-standard approach to establish such connections. Focusing on reaction networks of metabolites and the respective catalysts, algebraic procedures of coarse-graining are proposed as a natural tool to switch between multiple scales. In this regard, the joint and subsequent functions of single catalysts and of subnetworks on the reaction network are defined in sections 5.2-5.4. The set of the functions of all subnetworks forms a semigroup under composition. It is then demonstrated that such semigroups can be used to identify self-sustaining subnetworks (section 5.5). Finally, biologically meaningful congruences and the resulting coarse-graining procedures are defined and discussed (section 5.6).

Chapter 1

Investigated System and Applied Methods

This chapter introduces the tryptophan synthase enzyme as the main system under investigation in this thesis and the methods used to study it. In section 1.1, details on the structure and function of the enzyme are given. In section 1.2, approaches to model protein kinetics are discussed. Section 1.3 introduces the material on stochastic and information thermodynamics used in this thesis.

1.1 The Tryptophan Synthase Enzyme

The enzyme tryptophan synthase catalyzes the last two steps in the formation of L-tryptophan (in the following: tryptophan) from indole glycerol phosphate (IGP) and L-serine (in the following: serine). It is present only as a dimeric $\alpha_2\beta_2$ bienzyme complex with linear $\alpha\beta\beta\alpha$ alignment of the subunits. The α-subunit catalyzes the formation of indole and glyceraldehyde-3-phosphate (G3P) from IGP. Indole is then transferred through a 25 Å-long tunnel to the β-subunit, where it reacts with serine to form tryptophan (figure 1.1). To prevent loss or accumulation of the metabolite indole, the reactivity of both subunits is tightly coupled by allosteric interactions. Binding of both substrates IGP and serine triggers the closing of the α and β-subunits and thereby significantly enhances the rate of indole formation. Only after indole channeling to the β-site and reaction with serine is completed the subunits are opened and the product tryptophan and G3P released.

Tryptophan synthase has been extensively studied since 1946, when first indications for the biosynthesis of tryptophan from serine were given by Gunsalus [26]. Already in 1958 it was discovered that IGP and serine react to form tryptophan without releasing indole into the solution [27]. Since 1970, kinetic and structural studies performed by the groups of Michael F. Dunn (University of California, Riverside) and Ilme Schlichting (Max Planck Institute for Medical Research, Heidelberg) have created a vast amount of insights and data on tryptophan synthase. By the late 1990s the most important intermediates in the enzyme's cycle have been spectroscopically characterized and the reaction mechanism could be formulated. Since then research was focused on the understanding of the regulatory pathways synchronizing the α- and β-reactions.

A growing number of X-ray crystallographic structures of the wild-type enzyme and

mutants thereof with naturally appearing and model ligands has aided to identify the domains and residues responsible for catalysis and allosteric regulation [28]. In addition, several kinetic studies involving mutant enzymes and isotopically labeled substrates have been conducted to identify the rate determining reaction steps and the residues involved therein. In 2007, a further milestone was set by determining the X-ray crystallographic structure of tryptophan synthase in its closed and catalytically active conformation [29]. The historical development of research on tryptophan synthase and the interconnection of experimental results and their implications are reviewed in [19]. Articles that focus on structural [30] and kinetic [31] properties of tryptophan synthase are also available.

Higher organisms obtain the essential amino acid tryptophan through their diet, while bacteria, yeasts and molds have a tightly controlled mechanism for its synthesis regulated by the tryptophan operon. Hence, the elucidation of the mechanisms governing the enzyme's behavior is of interest in areas related to the medicine of infectious disease, plant defense and herbicide design.

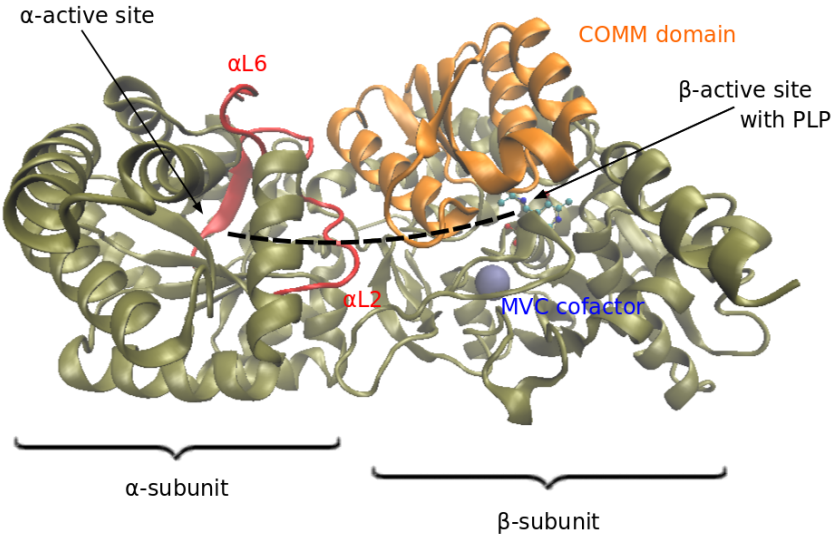


Figure 1.1: Structure of tryptophan synthase with its characteristic elements. The tunnel for indole channeling is represented by the dashed line. The COMM domain (orange) serves for allosteric information transfer between the subsites and prevents the escape of substances at the β-site in the closed conformation. The loops αL2 and αL6 (red) confer the allosteric communication at the α-site. In the closed conformation they prevent substrate exchange of the α-site with the enzyme environment. A ball and stick representation is used for the PLP cofactor at the β-site. PDB code: 2J9X.

1.1.1 Structural Features

The α-reaction

At the α-site of tryptophan synthase, indole-3-glycerol phosphate (IGP) is converted to indole and glycerol-3-phosphate (G3P) (figure 1.2). From X-ray crystallographic studies it is known that the α-subunit exists in at least two conformations termed as open and

closed states [32, 28]. The open state has a low catalytic activity on IGP cleavage and is structurally characterized by a disordered αL6 loop consisting of the residues α179-α193, which becomes ordered in the closed conformation and prevents the escape of indole into solution [33, 34, 32]. Concerning the reaction mechanism for aldolytic cleavage of IGP, two alternatives have to be taken into account. The first is a series of proton transfers involving αGlu49 and αAsp60 as acid-base catalysts and the second is a concerted one-step reaction. Considering the hydrophobic microenvironment of the active site, the latter mechanism seems to be more likely [35, 36, 29]. Using a specific α-site ligand, transition state analogues supporting the hypothesis of a concerted mechanism could be synthesized and analyzed crystallographically [37, 38].

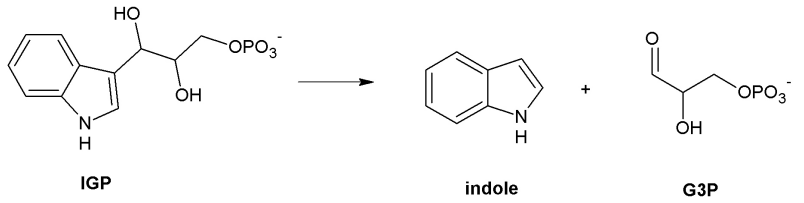


Figure 1.2: Transformation of IGP to indole and G3P catalyzed by the α-site of tryptophan synthase

The β-reaction

The β-subunit catalyzes the conversion of indole and serine to tryptophan (figure 1.3). In the initial state E(Ain), the cofactor pyridoxal phosphate (PLP) is bound to βLys87. It constitutes the main catalytic site for the complex reaction cycle by binding of the substrates through their amino groups as aldimines, germinal diamines and quinolines. So far, nine intermediates have been characterized by UV/Vis spectroscopy, X-ray crystallography and by reaction and comparison with substrate analogues [39, 40, 41]. The β-reaction is commonly divided in two stages. In stage I, the aminoacrylate E(A-A) is formed from the internal aldimine E(Ain) with serine with the germinal diamine E(GD_1), the external aldimine E(Aex_1) and the quinoline E(Q_1) appearing as intermediate states. In stage II, E(A-A) reacts with indole to give tryptophan and return to the enzyme's initial state E(Ain) via two quinolines E(Q_2) and E(Q_3), an external aldimine E(Aex_2) and a germinal diamine E(GD_2). As the first step of this stage, indole is channeled from the α-site to react with E(A-A). Like the α-subunit, the β-subunit can adopt at least two different conformations - an activated state with a closed conformation and an inactive open state. The catalytic cycles of the α- and β-sites are synchronized through a mechanism wherein conversion of E(Aex_1) to E(A-A), via E(Q_1), activates the α-site, whereas conversion of E(Q_3) to E(Aex_2) brings it back to the inactive open conformation [41, 42, 43]. In order to accommodate many different intermediates and thereby achieve reasonable reaction rates, the β-catalytic site possesses a certain structural flexibility, which is modulated by a monovalent cation (MVC) cofactor [44, 45, 46].

Mechanisms of Intersite Communication

Three levels of events comprise the allosteric communication in tryptophan synthase. These consist of loop motions at the α-site (loop αL2 with residues α53 to α60 and loop

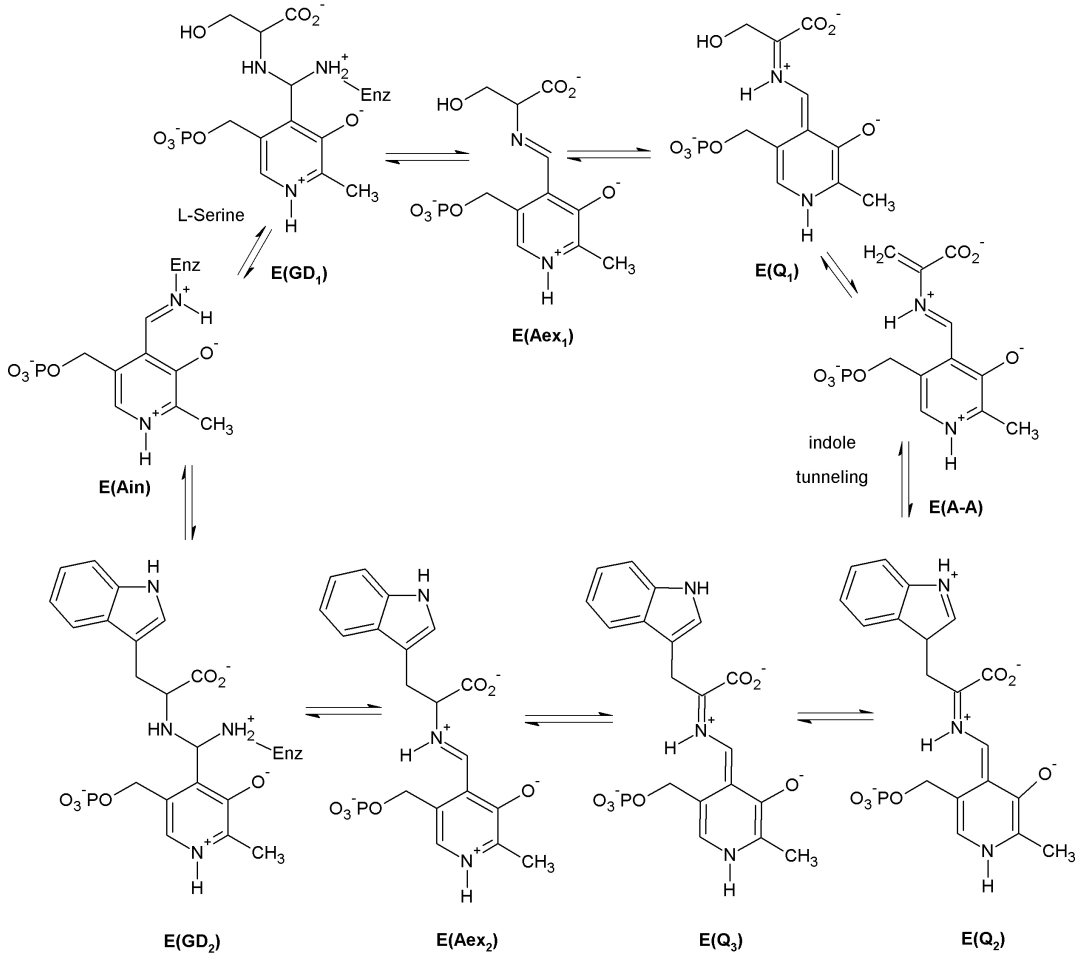


Figure 1.3: The β-reaction cycle catalyzed by tryptophan synthase. Serine reacts with the internal aldimine E(Ain) and is transformed to aminoacrylate E(A-A) under elimination of water. E(A-A) incorporates indole to yield the geminal diamine E(GD_2) via several intermediates, which releases tryptophan and returns to the initial state E(Ain).

αL6 with residues α179 to α193), motions of single residues extending over the bienzyme complex and motion of the COMM domain (residues β102 to β189). These movements are correlated, but the extent of concertion has yet to be established. The known communication mechanisms will be described in the above given order.

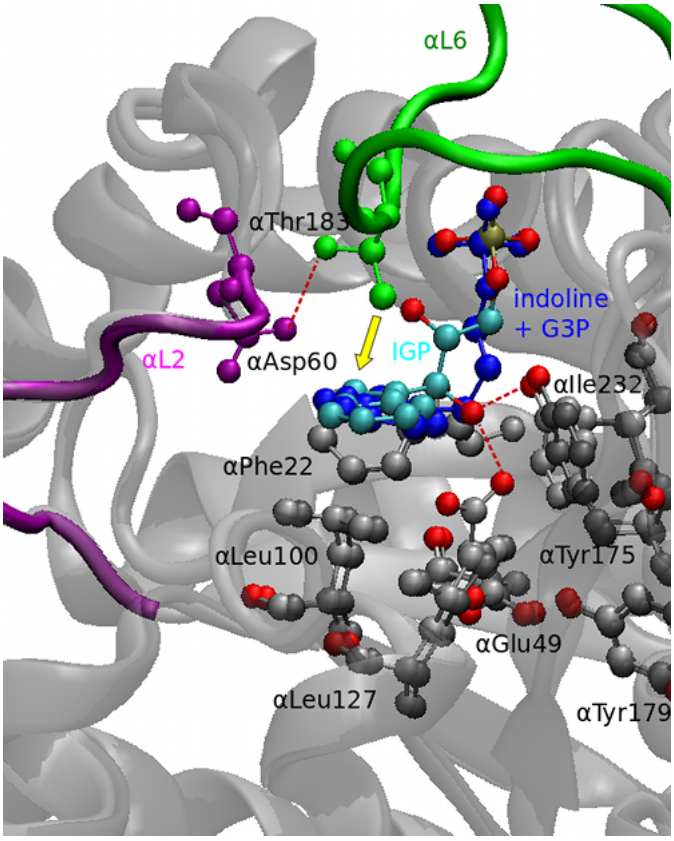


Figure 1.4: Conformational rearrangements in the α-subunit. The structures of an indoline-G3P adduct (dark gray, PDB code: 1QOP) with the IGP complex (light gray, PDB code: 2RHG) are compared. When the enzyme switches to the closed conformation, the loop αL6 (green) moves towards the substrate IGP. In the process, αThr183 gets pulled by αAsp60 through hydrogen bridge formation and pushes the substrate (yellow arrow). At this moment, IGP is able to interact with αGlu49 and αTyr175, which confer the concerted catalytic cleavage of IGP to G3P and indole. The residues αPhe22, αLeu100, αLeu127 and αIle232 form a suitable binding pocket for the product indole. The figure was rendered with VMD and modified with Inkscape.

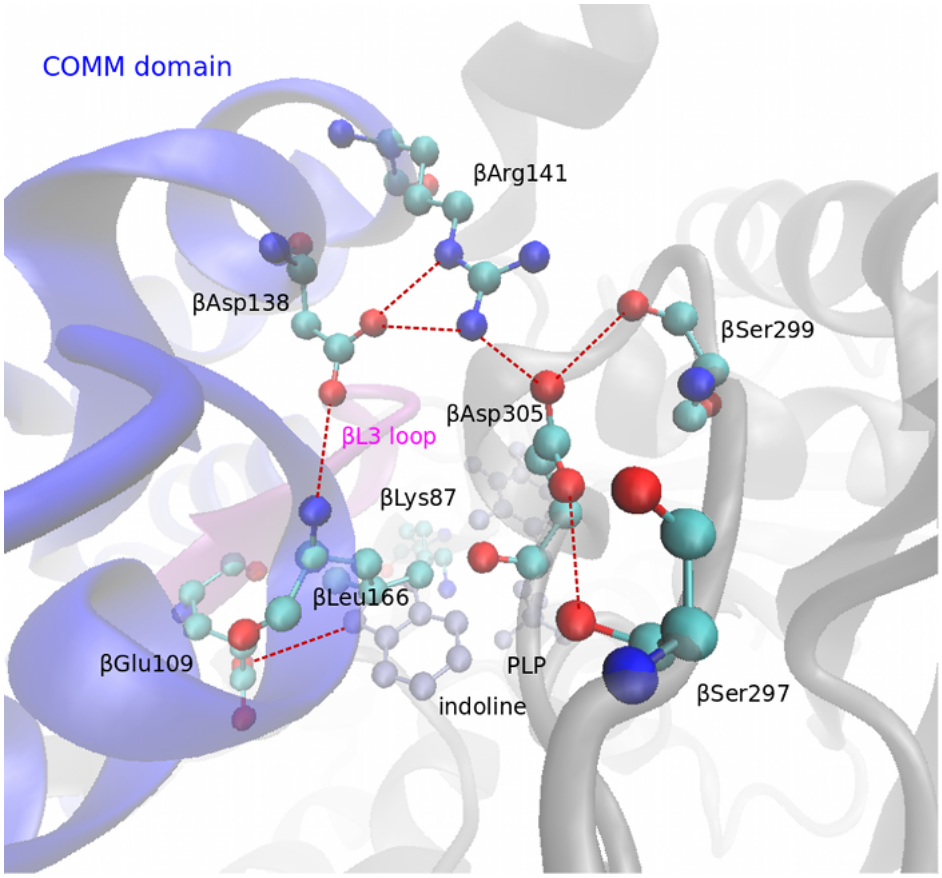


Figure 1.5: Hydrogen bonding network in the indoline derivative of the enzyme state E($Q_{2/3}$) (PDB code: 3CEP). When the enzyme adopts its closed conformation accompanied by the release of water at the reactive site, the residue βAsp305 rotates towards βArg141, which in turn moves 4 Å towards βAsp305. βGlu109 moves towards the substrate and forms a hydrogen bond with the indoline ring. The bonding network serves to stabilize certain intermediates in the closed conformation and is thought to prevent mass exchange with the environment. Hydrogen bonds are represented by dashed red lines. The figure was rendered with VMD and modified with Inkscape.

By using α-site ligand derivatives, it was possible to show that during the transition from the open to the closed conformation the loop αL2 moves towards αL6 and a crucial hydrogen bond is established between αThr183 on αL6 and αAsp60 on αL2 [29, 47]. αAsp60 then is orients so that it can stabilize charge developing during indole formation [37, 38] (figure 1.2). The residue αGlu49 is as well involved in proton transfer from C'-OH leading to the formation of indole via a push-pull mechanism. By X-ray crystallographic structures it has been shown to adopt two conformations: an inactive state with αGlu49 pointing away from the substrate [39] and the active conformation oriented towards the indole C'-OH group [48, 49]. This is assumed to be the most important interaction at the α-site for allosteric communication [30]. The structural details are shown in figure 1.4.

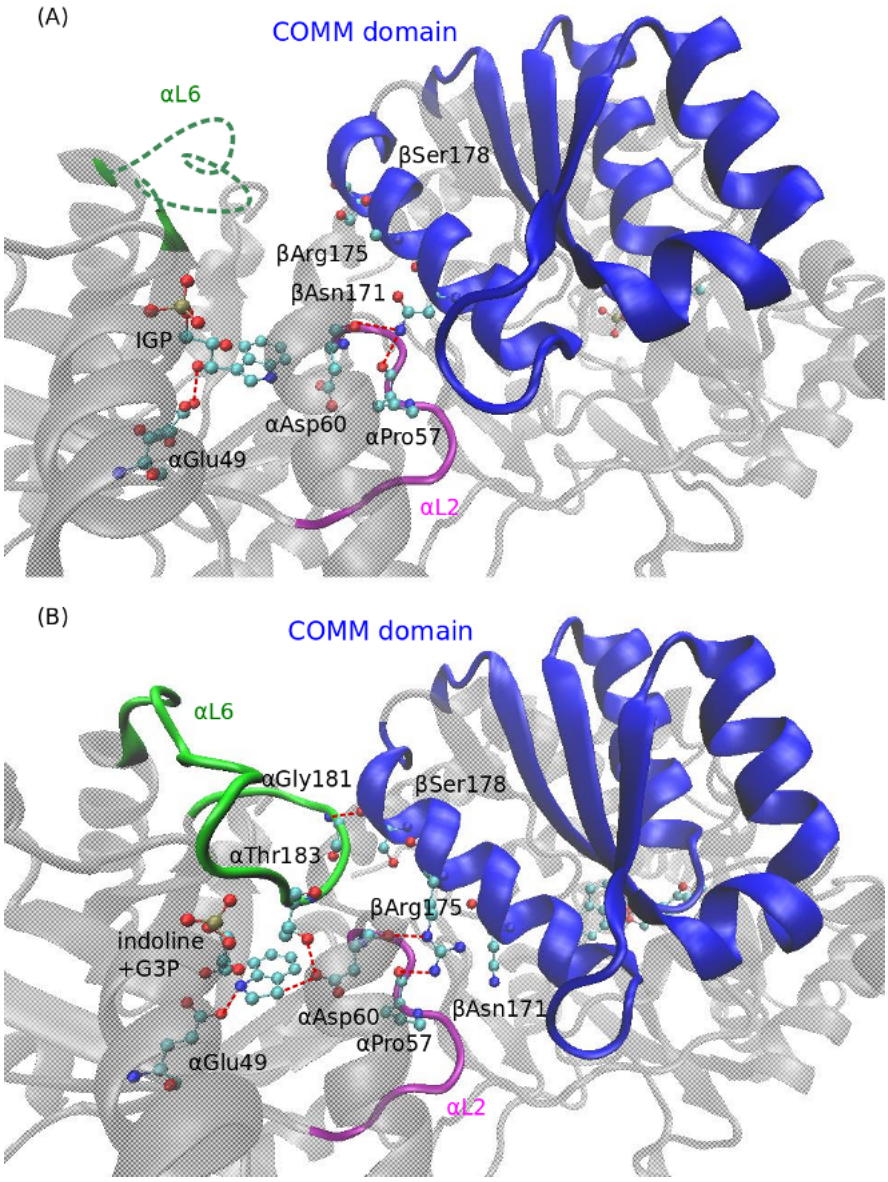


Figure 1.6: Comparison of the open and closed forms of tryptophan synthase at the interface between the α- and β-subsites. Red dashed lines denote hydrogen bonds. **(A)** Structure of the open state (IGP|Ain) (PDB code: 1QOQ). The open form is characterized by a disordered αL6 loop (green) and interactions of the αL2 loop with the COMM domain (blue) via hydrogen bond formation from αPro57 and αAsp60 to βAsn171. IGP is bound to αGlu49. **(B)** Structure of the indoline derivative of the closed state (G3P+indole|$Q_{2/3}$) (PDB code: 3CEP) The αL6 loop is now ordered and αThr183 interacts with αAsp60 and the substrate (compare with figure 1.4). In addition, αGlu181 forms a hydrogen bond to βSer178 on the COMM domain. The αL2 loop is closer to the substrate than in the open conformation thereby enabling interactions between αAsp60 and IGP. The COMM domain is displaced by one turn thus placing βArg175 in contact to αPro57. The figure was rendered with VMD and modified with Inkscape.

At the β-site, the β loop (residues β109 to β115) on the COMM domain confers a highly specific binding site for the substrate's and intermediates' carboxylate groups.

Surprisingly, the main conformational changes occur elsewhere: In the open E(Aex_1) structure, the carboxylate group of βAsp305 binds to the hydroxyl group of the serine moiety, thereby stabilizes the E(Aex_1) intermediate and prevents dehydroxylation by the acid-base catalytic βGly109 and βLys87 residues. Switching to the closed conformation leads to a movement of βArg141 by approximately 4 Å towards βAsp305. At this stage, the hydrogen bridge between βAsp305 and serine is broken and βAsp305 rotates about 100° [19]. This leads to an extended hydrogen bonding network between the residues βArg141, βAsp305, βSer297, βSer299, βAsp138, and βLeu166 [50, 28, 38, 48, 51, 52] (figures 1.6 and 1.5).

The mobile domain, which has been termed the COMM domain [50], consisting of the residues β102 to β189 is the key element in synchronization of α- and β-reactions. Its position defines the closed and open states of the β-subunit and couples to loops αL2 and αL6. In its open state the β-site is freely accessible from solution [39] while in the closed state the COMM domain moves towards the PLP cofactor closing the site and establishing interactions with other parts of the enzyme [38]. Within the COMM domain the helix βH6 is the main hub for intersite allosteric communication. In the open state, the residue βAsn171 on βH6 interacts with αAsp60 on αL2, which is part of the α-catalytic center [30]. When adopting the closed conformation, βArg175 interacts with αAsp60 and also αPro57. Moreover, hydrogen bridges are formed between βSer178 on βH6 and αGly181 on αL6 [53, 54] (figure 1.7).

The Monovalent Cation (MVC) Cofactor

In 1995, the group of Peracchi discovered that the tryptophan synthase enzyme utilizes a monovalent cation (MVC) cofactor [55]. It is bound to six carbonyl groups belonging to the residues βVal231, βGly232, βGly268, βLeu304, βPhe306, and βSer308, which form a loop around the cofactor [19]. The binding site is positioned 8 Å away from the β catalytic center [56]. Without the presence of the MVC cofactor, both the catalysis at the β-subsite and the allosteric communication are impaired. Removing the cofactor renders the aminoacrylate E(A-A) essentially unreactive towards indole [57]. Interestingly, the exact choice of the MVC species is rather robust towards size and charge density: Na^+, K^+, NH_4^+, Rb^+ and Cs^+ can serve as MVC cofactors [58, 44] and surprisingly also the large guanidinium ion [46, 59]. While the mechanistic influence of the cofactor on the allosteric communication has not yet been clearly worked out, modulation of the β reaction center has been clarified by analysis of crystal structures with different MVC cofactors. While the Cs^+-bound enzyme E(Cs^+) exhibits a binding pocket suited for indole and derivatives thereof, the pocket is too small in the Na^+-bound form E(Na^+)[42, 60, 56]. Consistently, the form E(Cs^+) favors the closed conformation and allows indole channeling and incorporation at the β-site and the form E(Na^+) favors the open conformation, where the formation of indole is kinetically hindered and thus a binding pocket for indole is not needed. In conclusion, the MVC cofactor is able to modulate the enzyme activity and to discriminate between the open and closed conformations. This is supported by the fact that for different cofactors different steady-state distributions of the respective enzymatic species have been measured [45, 61].

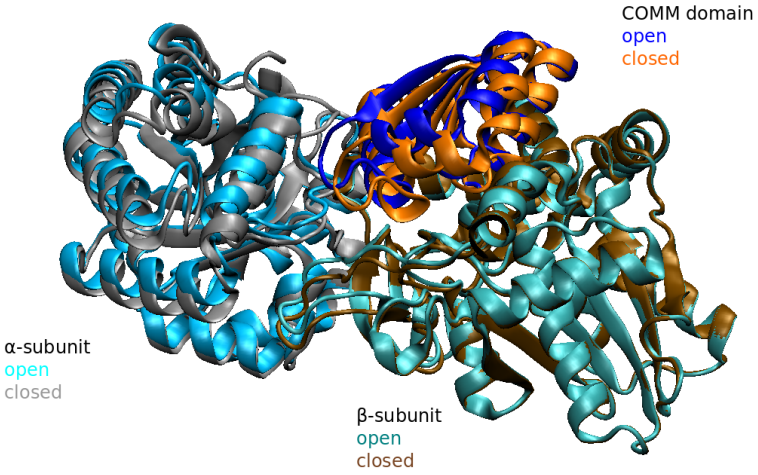


Figure 1.7: Superposition of the structures of open (PDB code: 1KFK) and closed (PDB code: 2J9X) conformations of tryptophan synthase. The COMM domain performs an extensive tilting motion, whereas the rest of the β-subunit does not change detectably. The α-subunit undergoes slight conformational changes. The figure was rendered with VMD and modified with Inkscape.

1.1.2 Kinetics of Tryptophan Synthase

The reaction cycle involving all known enzyme states is shown in figure 1.8 (with the labels from figures 1.2 and 1.3). Each subunit is represented by a chain and mutual regulations are indicated by colored arrows. The following allosteric interactions are highlighted in the literature [19, 31]

1. The state α-IGP has an activating effect on the formation of β-A-A: the reaction rate increases 9.7-fold. This result was obtained by Ngo *et al.* by using α-site ligands (ASL) that closely resemble the structure of IGP, but cannot be cleaved. The equilibrium distribution of the predominant β-species β-Aex$_1$ and β-A-A was then analyzed for the native enzyme with and without different ASL [29].

2. β-A-A in turn activates the formation of α-indole + G3P: the reaction rate increases 27.7-fold. This result was obtained by Brzovic *et al.* with similar methods as used by Ngo *et al.*. By binding serine analogues that could form β-A-A, but did not react further to the β-site, the rate of IGP cleavage could be measured and compared to rates with bound serine analogues that could not form β-A-A [42].

3. α-indole + G3P can only form when the enzyme is in the closed state. Therefore the β-site has to be in one of the following states: E(Q$_1$), E(A-A), E(Q$_2$) or E(Q$_3$) in order to enable the formation of α-indole + G3P.

4. In the closed conformation, the uptake and release of substrates and products is not possible. For the actual mechanism of the tryptophan synthase enzyme it has been

suggested that the states E(Q_1), E(A-A), E(Q_2) and E(Q_3) can exist in the open conformation [62]. Therefore, for these chemical states, two different conformational states - open and closed - have to be distinguished. In the former case, mass exchange with the environment is possible.

5. As discussed in section 1.1.1, the conversion IGP → G3P + indole most likely takes place as a concerted one-step reaction and no intermediate steps have to be taken into account.

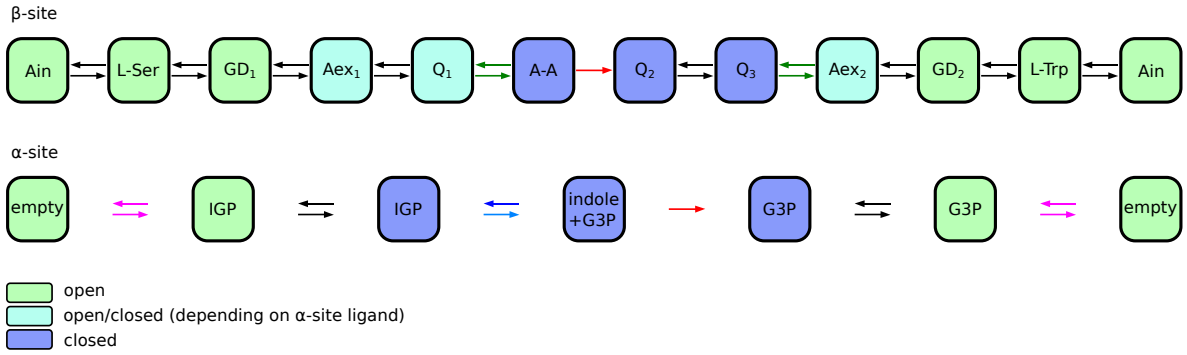


Figure 1.8: Allosteric interactions between the two subunits. The transitions empty ⇌ IGP and G3P ⇌ empty (magenta) in the α-site are blocked (i.e., the gate in the α-subunit is closed) in the states A-A, A-A + indole and Q_3 of the β-site. The transitions IGP ⇌ indole+G3P (light and dark blue) in the α-site are blocked in the states empty, Q_1, Aex_2 of the β-site. The rate of the transition IGP → indole+G3P (light blue) in the α-site is enhanced by a factor of 27.7 in the state A-A of the β-site. The transitions Q_1 ⇌ A-A and Q_3 ⇌ Aex_2 (green) in the β-site are blocked in the state empty of the α-site. The transition Q_1 → A-A (light green) in the β-site is enhanced by a factor of 9.7 in the state IGP of the α-site. The changes indole+G3P ⇌ G3P and A-A ⇌ indole+A-A (red) corresponding to indole channeling from the α- to the β-site occur simultaneously and represent a single stochastic transition.

A simplified scheme of the catalytic cycle of tryptophan synthase with several omitted states is displayed in figure 1.9. Here, the α-subunit is shown in green and the β-subunit in blue. The chemical states have the same notations as in figure 1.8. The catalytic cycle begins with the enzyme in the state where both sites are empty and the gates are open. Then, the substrate IGP binds to the α-subunit and serine to the β-subunit, where it is quickly converted to the serine quinoline intermediate Q_1. IGP activates the formation of the α-aminoacrylate A-A and the enzyme adopts the closed conformation, as schematically shown in figure 1.9b. In the state (IGP,A-A) where both gates are closed, A-A activates the cleavage of IGP to produce G3P and indole. Indole is then channeled to the β-site where it reacts with A-A to give the tryptophan quinoline intermediate Q_3 that is converted to tryptophan (Aex_2 is the external aldimine of tryptophan in the β-subunit). In the state (G3P,Aex_2) the gates open and the products tryptophan and G3P are released. Thus the enzyme returns to the initial conformation (empty,empty) and is ready to start the next cycle.

The kinetic rates for all transitions are given in section 2.2.

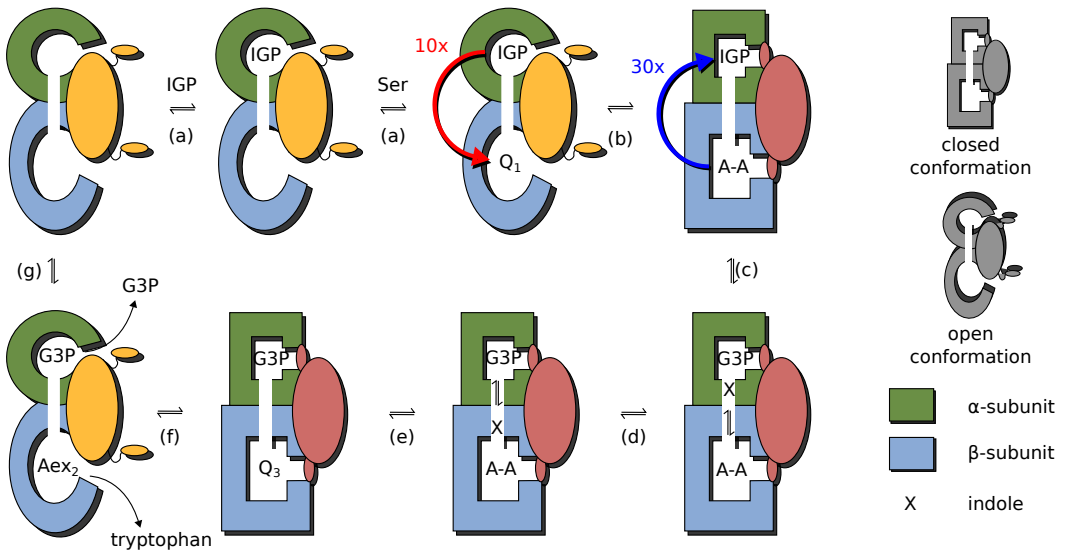


Figure 1.9: Schematic operation of tryptophan synthase. Operation of the machine: Once substrates are bound (a) at both catalytic sites, IGP activates (b) the formation of A-A and the enzyme adopts the closed conformation. A-A activates (c) the cleavage of IGP and indole is channeled (d) to the β-site where it reacts (e) with A-A to give Q_3. Q_3 undergoes (f) further transformations that return of the enzyme to the open conformation where tryptophan and G3P are released (g).

1.2 Protein Models and Protein Kinetics

There are several methods to model the structure, dynamics and kinetics of proteins. Such methods include quantum mechanics (QM), all-atom molecular mechanics (MM) or molecular dynamics (MD), hybrid QM/MM approaches, coarse-grained structural models such a Go models and elastic network models and phenomenological models with strongly reduced state spaces such as discrete Markov chains for chemical reactions, low-dimensional continuous parametrizations for conformational dynamics or a combination of both. The methods differ in the phenomena they are able to describe and in the time scales they are able to address. The most fundamental level is the description of a protein as a quantum mechanical system providing the full information on its electronic structure. All-atom molecular dynamics (MD) models contain full information on the coordinates of the nuclei, but take into account the electronic interactions via ad hoc potentials between groups of nuclei. In phenomenological models, qualitative or quantitative experimental data on the protein under consideration governs the choice of the variables in the model. Often the state space in such models is substantially reduced in comparison to MD models as many conformational and chemical states are not resolved, but treated as combined *coarse-grained* states. The time and length scales of the phenomena under investigation and the available experimental information determine the choice of the modeling approach.

Electronic processes in proteins take place on time scales of picoseconds, they are quantum chemical phenomena and have been modeled accordingly [63]. Examples of biologically relevant quantum mechanical processes are photon absorption in light harvesting complexes, substrate binding, proton and electron tunneling and chemical reactions catalyzed by enzymes. The light harvesting complexes photosystem I (PS I) and photosystem

II (PS II) play a main role in the transformation of the energy from absorbed photons into chemical energy and thus have been studied extensively. The absorption spectra of chlorophyll complexes in PS I and their dependence on the complex geometry have been determined by semiempirical methods [64]. Recently, the absorption spectrum of PS II was determined with *ab initio* methods [65]. Moreover, in the case of PS II, the pathway of electron absorption could be modeled. It involves 6 cofactors coupled to 4 charge-transfer states. The characteristic time scales were obtained as well [66]. The [$CaMn_3$(III)Mn(II)] cofactor of PS II catalyzes the splitting of water and production of oxygen; the mechanism of the reaction is still a topic of debate. The magnetic and electronic properties of the complex were calculated paving the way to a better understanding of the reaction mechanism [67]. In general, the electronic structure of metal cluster cofactors is important for the understanding of many biochemical processes, yet difficult to access. Another example are iron-sulfur clusters present in various classes of enzymes. Recently, it became possible to perform *ab initio* calculations of the energy landscape of [2Fe-2S] and [4Fe-4S] clusters without any fitting parameters [68]. Quantum chemical models have also been employed to determine binding energies of CO, NO and O_2 to heme molecules [69]. The study revealed a change in the magnetic structure of the Fe(II) center upon NO binding as compared to CO and O_2 ligands. Proton tunneling [70, 71] and electron tunneling [72, 73, 74, 75] pathways have been determined. Free energy barriers of chemical reactions in solution are accessible via quantum chemical methods [76]. There have also been attempts to model the dynamics of whole proteins using density functional theory [77, 78].

However, generally it is not possible to reach time scales relevant for the conformational dynamics of proteins with using quantum chemical models. A popular approach to retain the accurate description of electronic processes provided by quantum mechanics and to simultaneously study the conformational dynamics of a protein is the hybrid quantum mechanics/molecular mechanics (QM/MM) approach [79, 80, 81]. Thereby, the chemical reaction center is modeled as a quantum chemical system and the protein backbone by MM methods. For example, a QM/MM hybrid approach allowed to model the catalytic reaction of cAMP-dependent protein kinase [82]. The residues in the catalytic pocket responsible for a substantial reduction of the activation energy as well as residues that keep the substrates in an appropriate conformation were identified. As another example, a QM/MM model enabled the identification of a critical arginine residue in the catalytic mechanism of citrate synthase and allowed to study the interplay of conformational dynamics involving the arginine residue and catalytic activity [83]. Similarly, the coupling of vibrational excitations and catalytic activity in human purine nucleoside phosphorylase [84] and the interplay of conformational and electronic states in cytochrome C450 oxidation [85] could successfully be modeled. Hybrid methods also allow to determine acidity constants, redox potentials and solvation free energies of proteins using *ab initio* calculations [86].

The QM/MM hybrid methods can successfully take into account small-scale conformational motions at the catalytic site, but are not capable of reproducing domain motions in proteins as they take place on time scales of micro- to milliseconds. In many cases, insights into protein function can be gained without quantum chemical descriptions, but purely from the conformational dynamics of the protein [87]. All-atom molecular dynamics (MD) simulations trace the motions of all protein and solvent atoms using phenomenologically

adjusted force fields. MD simulations played a major role in the determination of the catalytic mechanism of F_1-ATPase. After the determination of the protein structures of the main chemical and conformational states of the catalytic cycle by protein crystallography, MD simulations have been used to interpolate between the structures in a biologically meaningful way and thereby provided a dynamical model of the functioning of F_1-ATPase [88, 89]. Moreover, the ATP binding affinities in the different conformational states of the F_1-ATPase β-subunits were determined using MD and an analysis of the thermodynamics of the simulated trajectories. This provided the solution to a dispute concerning the reaction mechanism [90]. Another example of the success of MD is the insight into the activity of Src tyrosine kinases, whose activated forms are known to be oncogenes [91]. Src kinases posses a catalytic domain, an SH2 peptide binding domain at the N-terminus of the catalytic domain and an SH3 binding domain at the C-terminus. In the inactive state, the SH2 and SH3 domains are tightly bound and block the entrance to the catalytic center [92]. Using MD simulations, it was possible to clarify the activation mechanism of the kinase: The catalytic domain possesses an activation segment that induces rearrangements in the SH2 domain and thereby weakens the SH2/SH3-interactions through long-range allosteric interactions. This leads to an increased accessibility of the catalytic center [93, 94].

The time scales accessible with molecular dynamics simulations are typically on the order of nanoseconds [95]. Using specifically designed computer architectures, a 1 millisecond trajectory was calculated for small proteins [96], breaking the previous record of a 10 microsecond trajectory [97] by a 100-fold. Yet, even such state of the art simulation techniques cannot reach the time scales of protein folding or large domain motions in molecular machines which often take place on the order of milliseconds and seconds [98, 99, 100]. To model such phenomena, coarse-grained molecular dynamics methods are available [101]. Thereby, groups of atoms, whole amino acid residues or even protein domains are grouped together to single particles and the dynamics is determined by potentials between such coarse-grained particles. The potentials can be introduced ad hoc, derived from all-atom potentials [102, 103], from statistical analysis of protein structure data [104] or adjusted to the native structure of the protein (Go models) [105, 106]. Coarse-grained molecular dynamics leads to a 10^3-fold [107] to 10^7-fold [108] speedup in computation time as compared to all-atom MD. A particularly attractive field for the application of structure-based models is protein folding [109]. Such models were used to generate a large amount of folding trajectories for different proteins allowing a statistical analysis of the folding pathways and generating new deep insights into the process of protein folding [110, 111, 112]. Protein dynamics around the native state can be studied, for example, with elastic network models [113, 114]. In these computationally very efficient models, all amino acid residues are replaced by single point particles and particles within a given cutoff range interact through harmonic potentials. Using such models, it was possible to simulate the whole catalytic cycle of HCV helicase [115], to study the allosteric interactions in myosin-V [116] and even to simulate global ribosome motions [117].

If the full structure of a protein is not available, it is possible to construct a state space from kinetic measurements and other experimental insights and to determine the transition rate constants between the states experimentally. The state space can consist of different chemical and conformational states [118]. The chemical state space is

usually finite and discrete corresponding to the space of chemical intermediates occurring in the catalytic cycle. If the conformational motions are faster than the chemical reactions, then the conformational states can be absorbed into the chemical states yielding a discrete state Markov model. For example, the motor protein kinesin has been modeled in [119] as a Markovian process on a discrete state space determined by the chemical states of both legs. Hereby, each leg can adopt three different states (empty, ADP-bound and ATP-bound) resulting in nine different states. If the conformational motions are slower than the chemical reaction, the conformational motions are described by a drift process on a low-dimensional manifold given by collective coordinates. An example is a model of F_1-ATPase, where the rotatory motion is characterized by a continuous coordinate and the chemical states of the protein are discrete corresponding to the bound ligands (empty, ADP-bound and ATP-bound) [120]. Other phenomenological models for F_1-ATPase [121, 122], kinesin [123, 119], myosin V , [124], dynein [125] and flagellar motors [126] have been constructed. Any protein model with discrete chemical states and Markovian transitions between them is a phenomenological model in this sense. Phenomenological models are well suited to study global aspects of proteins such as thermodynamic efficiency or the mechanochemical coupling in protein motors [127].

In principle, the modeling approaches with higher temporal and spatial resolution can be converted to models with lower resolution via coarse-graining. Thereby, certain subspaces of the state space are lumped together into coarse-grained states. If the dynamics within the coarse-grained states is much faster than the transitions between them, i.e. there is a separation of time scales, then a Markovian dynamics on the full state space transforms into a Markovian dynamics on the coarse-grained state space. For example, applying the Born-Oppenheimer approximation to a quantum mechanical description of a protein and integrating out the electronic degrees of freedom leads to a molecular dynamics model. Replacing the centers of mass of certain domains in the MD model and integrating out the fast atomic motions within such domains leads to coarse-grained models. The transformed dynamics is necessarily stochastic as the exact position within the coarse-grained states cannot be traced and the transitions between coarse-grained states occur at random with some given transition probability rates in discrete spaces or as a diffusive processes in continuous spaces. Even at the quantum mechanical level there are already sources of stochasticity in the dynamics due to the uncertainty relation. The stochasticity introduced through coarse-graining is, however, fundamentally different from quantum mechanical uncertainty, because it is not forced *a priori* by natural law.

1.3 Stochastic Thermodynamics

Classically, thermodynamics is applicable only to large systems with macroscopic state variables such as temperature, internal energy and entropy. The changes of the state variables are deterministic and can be associated with the quantities of work, heat and entropy production. In order for the variables to be well-defined, their fluctuations are required to be negligibly small.

Microscopic systems such as single proteins and mesoscopic systems such as reaction networks with low numbers of reactants are subject to large stochastic fluctuations and thus the classical theory of thermodynamics is not applicable to these systems. However,

it has become possible to assign thermodynamic quantities to such systems and to quantify the amount of work, entropy production and entropy flow for individual transitions and thus for stochastic trajectories. A historically pivotal point is the work of Schnakenberg, who generalized thermodynamic forces and fluxes to microscopic systems with fluctuating dynamics arbitrarily far from equilibrium [128]. In near-to-equilibrium situations, he recovered the Onsager reciprocity relations. The theory was extended in the subsequent decades to include a stochastic interpretation of energetics for driven systems and led to first-law-equalities [129]. Moreover, the discovery of stochastic violations of the second law [130] led to the formulation of fluctuation theorems [131, 132] that reveal a symmetry for the entropy production of a system at steady-state. Jarzynski proved a relation between the average work required to drive a system in a nonequilibrium regime and the free energies between the initial and final states [133]. This relation was refined by Crooks [134] and extended by others [135, 136]. These culminated efforts led to a thorough definition of a stochastic entropy and the second law [137, 138]. A further development has been the closely related field of information thermodynamics [139, 140]. Since the foundations of stochastic thermodynamics are formulated in terms of probabilistic processes, the hole machinery of information theory is at hand and enables investigations of measurement feedback and information transfer in bipartite systems [23]. Thorough and technical treatments of stochastic thermodynamics are available in a review article by Seifert [141] and a monograph by Sekimoto [142].

1.3.1 Stochastic Thermodynamics of Chemical Systems

Consider a Markov process on a discrete state space X. Denote the states of X by $x, x', ...$ and let $w_{x,x'}$ be the transition probability rate for a transition from the state x' to x. The probability to find the system in the state x at time t is denoted by $p(x;t)$. Its time evolution is given by the equation

$$\frac{d}{dt}p(x;t) = \sum_{x' \in X} [w_{x,x'}p(x';t) - w_{x',x}p(x;t)]. \tag{1.3.1}$$

This equation is known as a master equation. The time derivative of $p(x;t)$ depends only on $p(x;t)$, because the process is Markovian, i.e. memoryless. Using the probability fluxes $J_{x,x'}$ defined as

$$J_{x,x'} = w_{x,x'}p(x';t) - w_{x',x}p(x;t), \tag{1.3.2}$$

the master equation can be rewritten as

$$\frac{d}{dt}p(x;t) = \sum_{x' \in X} J_{x,x'}. \tag{1.3.3}$$

Any Markov process on a discrete state space X can be represented as a directed graph with vertex set X and edges from vertex x' to vertex x labeled by $w_{x,x'}$. Such graphs are called Markov networks.

The kinetics of a single protein molecule can be modeled as a Markov network. In this case, X is the space of chemical or conformational states and the transition probability

rates $w_{x,x'}$ are given by the zeroth order reaction rate constants for transitions not involving any additional reactants or derived from higher order reaction rate constants by fixing the concentrations of all additional reactants involved. It is crucial that the transitions between states are memoryless. As discussed in the previous section, a discrete state space can be obtained by the coarse-graining of conformational degrees of freedom and therefore a separation of time scales is necessary. This is the case when X is the state of chemical states of the protein and the chemical reactions are considerably slower than the conformational motions. The phenomenological models with discrete state spaces discussed in the previous section are examples of such Markov network models.

The master equation 1.3.1 is formally identical to a classical kinetic rate equation where $p(x;t)$ is replaced by the concentration of the chemical species x at time t. There is, however, a substantial difference between the two descriptions: Concentrations are macroscopic variables and the classical kinetic rate equation describe *deterministically* the evolution of these variables. A single enzyme therefore cannot be described by classical rate equations. At any given time, the single enzyme is in exactly one state $x \in X$ and jumps between the states according to the given transition probability rates $w_{x,x'}$. This is a *stochastic* process whose realizations are random walks on the corresponding Markov network. The probability distribution of the process, however, evolves deterministically according to the master equation. The description via a master equation contains more information than the corresponding classical rate equations. For example, the stochastic model of a single enzyme allows to determine the turnover time distribution, whereas the corresponding deterministic rate equations only yield the mean value of this distribution (see chapter 2).

A central quantity for stochastic processes is the Shannon entropy S. At time t it is given by

$$S(t) = -\sum_x p(x;t) \ln p(x;t). \tag{1.3.4}$$

Note that there alternative definitions of entropy for chemical reaction networks taking into account the statistical factors due to the indistinguishability of molecules of the same chemical species [143]. However, the most commonly used definition is the one given in equation 1.3.4. The time derivative of $S(t)$ is given by

$$\frac{d}{dt}S = \frac{1}{2}\sum_{x,x'} J_{x,x'} \ln \frac{p(x';t)}{p(x;t)}. \tag{1.3.5}$$

It can be split as

$$\frac{d}{dt}S = \sigma - h, \tag{1.3.6}$$

where

$$\sigma = \frac{1}{2}\sum_{x,x'} J_{x,x'} \ln \frac{w_{x,x'}p(x';t)}{w_{x',x}p(x;t)} \tag{1.3.7}$$

is the entropy production and

$$h = \frac{1}{2}\sum_{x,x'} J_{x,x'} \ln \frac{w_{x,x'}}{w_{x',x}} \tag{1.3.8}$$

is the entropy flow. The units of the entropy production and the entropy flow are [energy · temperature^{-1}· time^{-1}]. There is a difference between the Shannon entropy of an arbitrary stochastic process and the Shannon entropy of a physical process described by a master equation. In the latter case the Shannon entropy should be defined as $S(t) = -k_B \sum_x p(x;t) \ln p(x;t)$ (k_B is the Boltzmann constant) to allow a physical interpretation of the splitting $dS/dt = \sigma - h$. Therefore, throughout the text, the Shannon entropy and all quantities derived thereof such as σ and h will be given in units of k_B.

The entropy flow, the entropy production and the time derivative of the Shannon entropy are sums of contributions from single transitions $\sigma_{x,x'}$, $h_{x,x'}$ and $s_{x,x'}$:

$$\sigma = \frac{1}{2}\sum_{x,x'} \sigma_{x,x'}, \text{ with } \sigma_{x,x'} = J_{x,x'} \ln \frac{w_{x,x'} p(x';t)}{w_{x',x} p(x;t)}, \tag{1.3.9}$$

$$h = \frac{1}{2}\sum_{x,x'} h_{x,x'}, \text{ with } h_{x,x'} = J_{x,x'} \ln \frac{w_{x,x'}}{w_{x',x}}, \tag{1.3.10}$$

$$dS/dt = \frac{1}{2}\sum_{x,x'} s_{x,x'}, \text{ with } s_{x,x'} = J_{x,x'} \ln \frac{p(x';t)}{p(x;t)}. \tag{1.3.11}$$

The entropy production $\sigma_{x,x'}$ is the product of the probability flux $J_{x,x'}$ with the generalized force $\ln(w_{x,x'}p(x';t)/w_{x',x}p(x;t))$ [128]. This is based on the expression for the entropy production in phenomenological thermodynamics. This connection is discussed in appendix A. By Jensen's inequality for convex functions, the entropy production is always nonnegative. It vanishes only under equilibrium conditions.

In an equilibrium state with equilibrium probability distribution $p_{eq}(x;t)$, the principle of microscopic reversibility imposes the vanishing of all fluxes $J_{x,x'}$, i.e. the transitions from x to x' have the same probability to occur as transitions from x' to x. This implies

$$w_{x,x'} p_{eq}(x';t) = w_{x',x} p_{eq}(x;t) \tag{1.3.12}$$

for all pairs of states x, x' by equation 1.3.2. This leads to the condition on the quotient $w_{x,x'}/w_{x',x}$ known as detailed balance

$$\frac{w_{x,x'}}{w_{x',x}} = \exp\left(\frac{F(x') - F(x)}{k_B T}\right), \tag{1.3.13}$$

where $F(x)$ is the free energy of the state x, k_B is the Boltzmann constant and T is the system's temperature. Detailed balance gives an interpretation of the entropy flow $h_{x,x'} = J_{x,x'} \ln(w_{x,x'}/w_{x',x})$. Using equation 1.3.13 one can write

$$h_{x,x'} = \frac{J_{x,x'}(F(x') - F(x))}{k_B T} \tag{1.3.14}$$

and thus $h_{x,x'}$ is the heat flux from the system to the environment for the transition between x and x'.

The rates $w_{x,x'}$ are constants and therefore the condition 1.3.13 holds not only at equilibrium, but for any probability distribution. For any cycle (i.e. a closed path) Γ on the Markov network, equation 1.3.13 implies

$$\prod_{\Gamma} \frac{w_{x,x'}}{w_{x',x}} = \exp\left(\sum_{\Gamma} \frac{F(x') - F(x)}{k_B T}\right) = 1. \tag{1.3.15}$$

More generally, transitions can be coupled to reservoirs, such as chemical reservoirs if the respective transition involves the release or binding of some chemical or thermal reservoirs that correspond to cooling or heating of the system in the respective transition. Then the equation of detailed balance needs to be modified as

$$\frac{w_{x,x'}}{w_{x',x}} = \exp\left(\frac{F(x') - F(x) + F_{x,x'}}{k_B T}\right), \tag{1.3.16}$$

where $F_{x,x'}$ is a contribution due to the coupling to a reservoir. For coupling to a chemical reservoir, $F_{x,x'}$ corresponds to the sum of Gibbs free energies of chemicals supplied by the reservoir. $F_{x,x'}$ can also be a mixed term due to coupling to several reservoirs. Schnakenberg established a relationship between the $F_{x,x'}$ terms due to local coupling and the thermodynamic forces they create [128]. Multiplying the quotients $w_{x,x'}/w_{x',x}$ over a cycle Γ in the network gives

$$\prod_{\Gamma} \frac{w_{x,x'}}{w_{x',x}} = \exp\left(\sum_{\Gamma} \frac{F(x') - F(x) + F_{x,x'}}{k_B T}\right) = \exp\left(\frac{1}{k_B T} \sum_{\Gamma} F_{x,x'}\right). \tag{1.3.17}$$

where the second equality is obtained from equation 1.3.15. The force $F_\Gamma = \sum_\Gamma F_{x,x'}$ corresponds to the macroscopic forces driving the cycle (see appendix A).

Having defined the entropy and free energy for a Markov network with arbitrary probability distribtion, other state variables can be defined under the same conditions using the classical thermodynamical identities as they are required to coincide with the classically defined variables in equilibrium. For example, the internal energy U of the system, defined as the expectation value of the energy $U = \langle E \rangle$, is related to the free energy via $F = U - TS$. Then the non-equilibrium free energy is

$$F = \langle E \rangle + T \langle \ln p(x) \rangle, \tag{1.3.18}$$

where $-\langle \ln p(x) \rangle$ is the Shannon entropy from equation 1.3.4. The other state variables are defined analogously.

An important field for applications of stochastic thermodynamics is provided by biochemical reaction networks [144, 143, 145, 146, 121, 147, 148, 149, 150, 151, 152]. All living systems operate far from equilibrium and and reactants are often present in small numbers in biological cells. At the level of single macromolecules, protein motors use chemical potential gradients to perform work in a strongly fluctuating environment. To

investigate these small biochemical systems from a thermodynamic point of view, methods from stochastic thermodynamics have been widely used [149, 150, 151, 152]. Stochastic thermodynamics has provided a general description of the coupling of chemical reservoirs to the work extraction by molecular motors [153, 118, 154, 155] and to particular motors like the F_1-ATPase [121, 147, 148, 156, 122] and walkers such as kinesins and myosins [157, 158, 119, 159, 160]. Moreover, the theory has been experimentally confirmed by the application of nonequilibrium methods to precise determinations of free energy landscapes of biomolecules [161, 162, 163].

1.3.2 Information Thermodynamics

The Shannon entropy plays a central role in the study of the thermodynamics of stochastic processes. However, the information theoretic character of the Shannon entropy of a Markov network has not been discussed so far. It becomes clearly visible when two systems are coupled by correlations as will be illustrated now. Correlations between a system X and a measurement device M are created, for example, during measurement processes. The information gained through measurement can then be used to extract additional work from the system. For the sake of readability, the time-dependence of all quantities is not written out explicitly in this section. In other words, the measurement changes the probability distribution $p(x)$ on X to a conditional probability distribution $p(x|m)$ depending on the outcome m (a state of M) after measurement. This changes the Shannon entropy $S(X)$ of X to a conditional entropy $S(X|M)$ given by

$$S(X|M) = \sum_x \sum_m p(m)p(x|m) \ln p(x|m) \tag{1.3.19}$$

and changes the free energy $F = \langle E \rangle + TS(X)$ (equation 1.3.18) to $F = \langle E \rangle + TS(X|M)$ by an amount of $T(S(X) - S(X|M))$. The quantity $S(X) - S(X|M)$ is known as mutual information $I(X, M)$. Using the definition of conditional probabilities and $p(x) = \sum_m p(m)p(x|m)$, it is more conveniently written as

$$I(X, M) = \sum_x \sum_m p(x, m) \ln \frac{p(x, m)}{p(x)p(m)}, \tag{1.3.20}$$

where $p(x, m)$ is the joint probability distribution of X and M, i.e. the probability to find X and M in the states x and m at the same time. Equation 1.3.20 shows that the mutual information is symmetric in its arguments, i.e. $I(X, M) = I(M, X)$. Moreover, equation 1.3.20 shows that the mutual information depends on the correlations established by the measurement. If the systems remain uncorrelated, i.e. if $p(x, m) = p(x)p(m)$, then $I(X, M)$ vanishes, $S(X) = S(X|M)$ and no new information is obtained through the measurement. If, however, $p(x, m) \neq p(x)p(m)$, then $I(X, M)$ is strictly positive and the entropy $S(X|M)$ is lower than $S(X)$ leading to a higher free energy of X enabling X to perform more work. This surplus of free energy can be extracted through appropriate feedback protocols [164]. The resetting of the measurement device is achieved by destroying the correlations between X and M and thus requires at least the amount $I(X, M)$ of work. More details, the application to complete measurement-feedback-reset cycles and the confirmation of Landauer's principle within this framework can be found in [164].

For the rest of this chapter, the idea sketched above is applied to coupled systems with continuous information exchange following [22, 23, 24]. Let A and B be two systems with discrete and finite state spaces and let $A \times B$ the joint system with the corresponding product state space, i.e. with states (a, b), where a and b are states of A and B, respectively. A and B will be referred to as subsystems with the respective marginal probability distributions $p_A(a) = \sum_b p(a, b)$ and $p_B(b) = \sum_a p(a, b)$. Assume that $A \times B$ has Markovian dynamics described by a master equation

$$\frac{d}{dt}p(a, b; t) = \sum_{a',b'}[w^{b,b'}_{a,a'}p(a', b'; t) - w^{b',b}_{a',a}p(a, b; t)], \tag{1.3.21}$$

where $w^{b,b'}_{a,a'}$ is the transition probability rate from (a', b') to (a, b). Assume further that A and B do not undergo simultaneous transitions, but can have an effect on each others transition rates, i.e. the rate of a transition from a' to a depends on the current state b and vice versa. This means that the transition probability rates can be rewritten as

$$w^{b,b'}_{a,a'} = \begin{cases} w^{b,b'}_{a} & \text{if } a = a' \\ w^{b}_{a,a'} & \text{if } b = b' \\ 0 & \text{if } a \neq a' \text{ and } b \neq b' \end{cases} \tag{1.3.22}$$

and the fluxes can be written as

$$J^{b}_{a,a'} = w^{b,b'}_{a,a'}p(a', b'; t) - w^{b',b}_{a',a}p(a, b; t) \text{ if } b = b', \tag{1.3.23}$$

$$J^{b,b'}_{a} = w^{b,b'}_{a,a'}p(a', b'; t) - w^{b',b}_{a',a}p(a, b; t) \text{ if } a = a'. \tag{1.3.24}$$

Such a system is called bipartite as the corresponding Markov network is a bipartite network. Physically, one can think of this system as two subsystems A and B that continuously perform measurements on each other, while each of the subsystems also has an internal dynamics. This affects the *apparent* entropy productions σ^A and σ^B in both subsystems. These are defined as follows: Suppose that the subsystem A is observed without the knowledge of the subsystem B, i.e. there is no access to the joint probability distribution $p(a, b)$, but only to $p_A(a)$. The apparent entropy production σ^A assigned to the subsystem A is

$$\sigma^A = \frac{1}{2}\sum_{a,a',b} J^{b}_{a,a'} \ln \frac{w^{b}_{a,a'}p_A(a')}{w^{b}_{a',a}p_A(a)}. \tag{1.3.25}$$

Similarly,

$$\sigma^B = \frac{1}{2}\sum_{a,b,b'} J^{b,b'}_{a} \ln \frac{w^{b,b'}_{a}p_B(b')}{w^{b',b}_{a}p_B(b)}. \tag{1.3.26}$$

The time derivative $dI(A, B)/dt$ of the mutual information

$$I(A, B) = \sum_{a,b} p(a, b) \ln \frac{p(a, b)}{p_A(a)p_B(b)} \tag{1.3.27}$$

can be split into two terms

$$\frac{d}{dt}I(A,B) = i_A + i_B \tag{1.3.28}$$

with

$$i_A = \frac{1}{2}\sum_{a,a'}\sum_{b} J^b_{a,a'} \ln \frac{p_B(b|a)}{p_B(b|a')} \tag{1.3.29}$$

$$i_B = \frac{1}{2}\sum_{b,b'}\sum_{a} J^{b,b'}_{a} \ln \frac{p_A(a|b)}{p_A(a|b')}. \tag{1.3.30}$$

As shown in detail in chapter 4 for general Markov networks, the information fluxes i_A and i_B and the apparent entropy productions σ^A and σ^B obey the *second-law-like inequalities*

$$\sigma^A - i^A \geq 0, \tag{1.3.31}$$

$$\sigma^B - i^B \geq 0. \tag{1.3.32}$$

Moreover, $dI(A,B)/dt$ vanishes in a steady-state and thus $i^A = -i^B$. This means that the mutual measurement process can change the apparent entropy production in the system A by i^A. In particular, σ^A can be negative, it is only bound by $\sigma^A \geq i^A$. Such a reduction in entropy production must be compensated by the a flow of information from the system B. The splitting 1.3.28 of $dI(A,B)/dt$ and the inequalities 1.3.31 and 1.3.32 were derived by Horowitz and Esposito [164]. In chapter 4, the theory is generalized to arbitrary Markov networks and applied to the tryptophan synthase enzyme.

The union of information theory with stochastic thermodynamics is by now on a solid foundation. It has been recognized that processes driven by thermal or chemical gradients are formally treated in the same way as processes driven by information. Seifert and Barato have made the notion of "information reservoirs" precise [165] and Mandal and Jarzynski have provided an example of a process that extracts work from information stored in a linear array of zeros and ones [166]. Fluctuation theorems were proven for processes with information transfer [167].

Chapter 2

Markov Network Model

In this chapter, the single-molecule Markov network model of tryptophan synthase is constructed. In section 2.1, the previous kinetic models of tryptophan synthase using classical chemical rate equations are briefly reviewed. It is concluded that in the case of tryptophan synthase, it is more natural to use a single-molecule model than models valid for a homogeneous mixture of many copies of the enzyme. The experimental data available in the literature is presented in section 2.2 and analyzed in section 2.3. The Markov network model is constructed in section 2.4 and the results such as the turnover time distribution and quantifications of correlations and synchronization are presented in section 2.5, followed by a discussion in section 2.6. This work has been published in [21].

2.1 Previous Kinetic Models

Kinetic models for tryptophan synthase have been proposed in almost all the kinetic studies referred to throughout the text. In most of the cases, not the whole reaction cycle, but only the reactions under experimental investigation were modeled. Classical kinetic rate equations were used to deduce the respective rate constants (see section 2.2). To the author's knowledge, only the models [168] by Lane and Kirschner (1983) and [169] by Anderson, Miles and Johnson (1991) model of the full cycle of either one subunit or of the whole enzyme. Both models are based on classical kinetic rate equations. They have the drawback that, in effect, the two catalytic centers in the same enzyme are treated as two different and statistically independent chemical species such that both the allosteric interactions and the indole channeling take place between the ensembles of chemical species and not within a single enzyme. Thus, in such models, correlations within a single enzyme could not be considered.

Lane and Kirschner provided a detailed model [168] for the catalytic cycle of the β-subsite using only the β_2 homodimeric enzyme. They have taken into account the following chemical species: E(Aex_1), E(Q_1), E(A-A), E(Q_2), E(Q_3) and E(Aex_2), which were the only species that detectably accumulated during the experiment. In the experiments, the α-subunit was not needed to complete the enzymatic cycle as indole was added to the reaction mixture. The influence of the α-subunit in the native $\alpha_2\beta_2$ form was discussed qualitatively in terms of allosteric regulations, but not incorporated into the model. Therefore, the model was not capable of describing the whole enzymatic cycle.

The reaction scheme used in the kinetic study by Anderson *et al.* [169] is shown in figure 2.1. While the reaction at the α-site was taken into account completely, the reaction mechanism at the β-site was oversimplified. Some chemical states, which were known at the time of model formulation, were neglected. This was not justified by Anderson *et al.* The classical kinetic description has been used to model the enzyme kinetics. Thereby, the α- and β-subunits were treated as two separate enzyme species. Rate equations were formulated for the concentrations of different chemical states of these two species. Even the rate of indole incorporation into the β-subunit was given by a second-order rate constant, which was dependent on the concentration of indole.

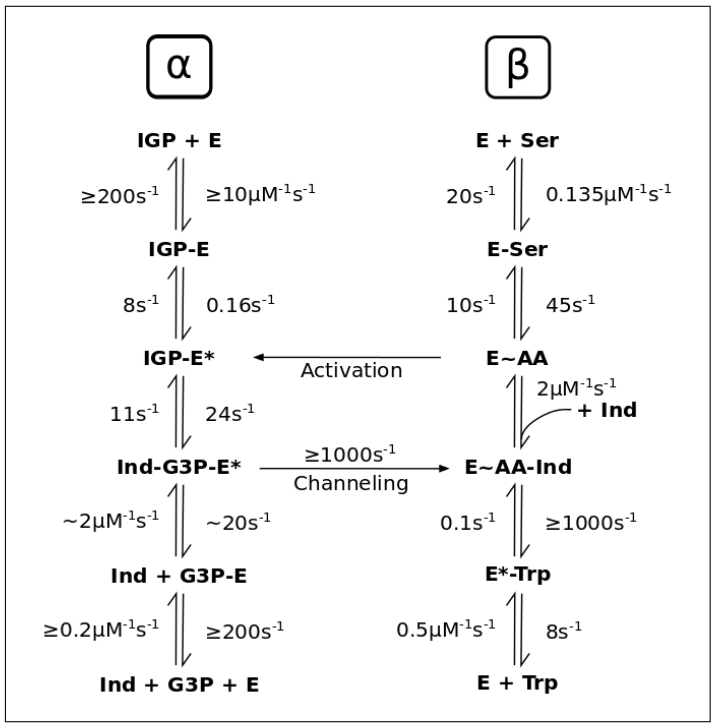


Figure 2.1: Scheme of the kinetic model for tryptophan synthase used by Anderson *et al.* in 1991 [169]. The figure is redrawn from the original publication with Inkscape.

As stressed in the review article by M. Dunn [19], the outstanding feature of tryptophan synthase is presence of strong correlations and synchronization in the states of the α- and β-subunits. During the catalytic cycle, the two subunits of one single enzyme molecule communicate by allosteric interactions and thereby stay in phase. The strongest correlations possible are present during indole diffusion through the tunnel: Both subunits have to take a specific state and only then the indole channeling can take place. Such strong correlations cannot be described by the classical model formulated by Anderson *et al.* [169].

The single-molecule model of tryptophan synthase constructed in this chapter can take into account the indole channeling and mutual allosteric interactions between the two catalytic sites. It is necessarily a stochastic model (see section 1.2) and in this case it is formulated in terms of a discrete state Markov network. For the network construction, both the important chemical species and transition rates between them need to be identified. This is done in the following two sections.

2.2 Kinetic Data

The kinetic data presented in this section was obtained by the groups of Dunn, Woehl, Schlichting, Anderson and Lane with rapid-scanning stopped-flow (RSSF) methods. RSSF experiments allow to measure transient species in fast chemical reactions with half-lives as low as a few milliseconds [170]. The reactant solutions are pushed from syringes into a reaction chamber where the mixing is almost instantaneous due to the small volume of the chamber. After mixing, reactant concentrations are measured by spectroscopic methods. The flow is stopped to increase the reaction time in the chamber and data is collected for the appropriate interval of reaction times to give the dependence of reactant concentrations on the elapsed reaction time. The kinetic parameters are then fitted numerically to the measured data. Note that a reaction mechanism needs to be proposed prior to the numerical analysis and that the analysis yields the rate constants of the proposed mechanism without any conclusions abouts its validity. To test the validity, simulations with the obtained rate constants are carried out and the results are compared to the experimental data. In the case of tryptophan synthase kinetics, the reactions at the β-site could be observed with RSSF techniques due to the flourescent PLP cofactor whose spectrum is sensitive to the different chemical species shown in figure 1.8 and thus allows to distinguish them by flourescence measurements. The data analysis was performed by the respective researchers using the KINSIM package [171].

Experimental studies of tryptophan synthase kinetics have covered many different aspects such as binding and release of substrates and products [172, 173], indole channeling [174, 175, 169], allosteric interactions [176, 177, 42, 178, 50, 35, 33, 29, 28], reaction of indole at the β-site after channeling [43], the reaction mechanism at the β-site with indole as a substrate [168, 179] or the effects of monovalent cation (MVC) cofactor [55, 58, 44, 45, 56, 46, 59, 61, 57]. Enzymes from both *Escherichia Coli* and *Salmonella Typhimurium* have been used. There were variations of pH-values, temperature, buffer solutions and salt concentration in these publications.

Data on the strength of allosteric interactions between the α- and β-subunits is taken from [29] and [42]:
Activation of the β-site: The state α-IGP has an activating effect on the formation of β-A-A: the reaction rate increases 9.7-fold. This result was obtained by Ngo *et al.* by using α-site ligands (ASL) that closely resemble the structure of IGP, but cannot be cleaved. The equilibrium distribution of the predominant β-species β-Aex_1 and β-A-A was then analyzed for the native enzyme with and without different ASL [29].
Activation of the α-site: β-A-A in turn activates the formation of α-indole + G3P: the reaction rate increases 27.7-fold. By binding serine analogues that could form β-A-A, but did not react further to the β-site, the rate of IGP cleavage could be measured and compared to rates with bound serine analogues that could not form β-A-A [42].

The kinetic rate constants for the reaction scheme in figure 1.8 are given in tables 2.1 and 2.2. They have been gathered from publications which focus on the whole catalytic cycle of tryptophan synthase [169], on the cycle of the β-subunit [168, 173] and on the influence of MVC on the reaction rates [44, 45].

Turnover rates for the whole enzymatic cycle have been determined as 3,29 s^{-1} [180], 5,0 s^{-1} [172], 3,8 s^{-1} and 4,6 s^{-1} [168] and 3,35 s^{-1} [42].

Reaction	Conformation	Conditions	Rate k	Source
IGP + TS → IGP-TS	open		$\geq 10\mu M^{-1}s^{-1}$	[169]
IGP-TS → IGP + TS	open		$\geq 200s^{-1}$	[169]
IGP-TS → IGP-TS*	unknown		$0.16s^{-1}$	[169]
IGP-TS* → IGP-TS	unknown		$8s^{-1}$	[169]
IGP-TS* → Ind-G3P-TS*	closed		$24s^{-1}$	[169]
Ind-G3P-TS* → IGP-TS*	closed		$11s^{-1}$	[169]
indole channeling	closed		$\geq 1000s^{-1}$	[169]
Ind-G3P-TS* → Ind + G3P-TS	open	loss of indole into solution	$20s^{-1}$	[169]
Ind + G3P-TS → Ind-G3P-TS*	open	indole uptake from solution	$2\mu M^{-1}s^{-1}$	[169]
G3P-TS → G3P + TS	open		$\geq 200s^{-1}$	[169]
G3P + TS → G3P-TS	open		$\geq 0.2\mu M^{-1}s^{-1}$	[169]

Table 2.1: α-Reaction: kinetic rate constants. The results from [169] were obtained using KINSIM. Abbreviations: TS: tryptophan synthase, KINSIM: kinetic simulation program. [171].

Reaction	Conformation	Conditions	Rate k	Source
TS + Ser → TS-Ser	open		$0.135\mu M^{-1}s^{-1}$	[169]
TS-Ser → TS + Ser	open		$20s^{-1}$	[169]
TS + Ser ⇌ TS-Ser	open	depends on NaCl concentration	$K = 0.07mM^{-1}$	[44]
TS + Ser → E(Aex_1)	open	no α-site ligand	$7.5 \cdot 10^4 M^{-1}s^{-1}$	[168]
TS + Ser → E(Aex_1)	open	with IPP	$7.5 \cdot 10^4 M^{-1}s^{-1}$	[168]
TS + Ser ⇌ E(Q_1)	open	pH = 7.6, no α-site ligand	$K = 0.72mM$	[173]
TS + Ser ⇌ E(Q_1)	open	pH = 6.4, no α-site ligand	$K = 2.54mM$	[173]
TS + Ser → E(Q_1)	open	pH = 7.6, no α-site ligand	$150mMs^{-1}$	[173]
TS + Ser → E(Q_1)	open	pH = 6.4, no α-site ligand	$45mMs^{-1}$	[173]
E(Q_1) → TS + Ser	open	pH = 7.6, no α-site ligand	$109s^{-1}$	[173]
E(Q_1) → TS + Ser	open	pH = 6.4, no α-site ligand	$113s^{-1}$	[173]
TS-Ser → E(A-A)	not specified		$45s^{-1}$	[169]
TS-Ser → E(Aex_1)	not specified	depends on [NaCl]	$1390s^{-1}$	[44]
E(Aex_1) → TS-Ser	not specified	depends on [NaCl]	$23s^{-1}$	[44]
E(Aex_1) → TS + Ser	open	pH = 7.6, no α-site ligand	$500s^{-1}$	[168]
E(Aex_1) → TS + Ser	open	pH = 6.4, no α-site ligand	$500s^{-1}$	[168]

$E(Aex_1) \rightarrow TS + Ser$	open	pH = 7.6, with IPP	$450s^{-1}$	[168]
$E(Aex_1) \rightarrow E(Q_1)$	not specified	no α-site ligand	$300s^{-1}$	[168]
$E(Q_1) \rightarrow E(Aex_1)$	not specified	no α-site ligand	$80s^{-1}$	[168]
$E(Aex_1) \rightarrow E(Q_1)$	not specified	with IPP	$267s^{-1}$	[168]
$E(Q_1) \rightarrow E(Aex_1)$	not specified	with IPP	$120s^{-1}$	[168]
$E(Aex_1) \rightarrow E(A\text{-}A)$	not specified		$16.7s^{-1}$	[44]
$E(A\text{-}A) \rightarrow E(Aex_1)$	not specified		$5.5s^{-1}$	[44]
$E(Q_1) \rightarrow E(A\text{-}A)$	not specified	no α-site ligand	$50s^{-1}$	[168]
$E(A\text{-}A) \rightarrow E(Q_1)$	not specified	no α-site ligand	$3s^{-1}$	[168]
$E(Q_1) \rightarrow E(A\text{-}A)$	not specified	no α-site ligand, alternative pathway	$13.3s^{-1}$	[168]
$E(A\text{-}A) \rightarrow E(Q_1)$	not specified	no α-site ligand, alternative pathway	$0.8s^{-1}$	[168]
$E(Q_1) \rightarrow E(A\text{-}A)$	not specified	with IPP	$15s^{-1}$	[168]
$E(A\text{-}A) \rightarrow E(Q_1)$	not specified	with IPP	$0.1s^{-1}$	[168]
$E(Q_1) \rightarrow E(A\text{-}A)$	open	pH = 6.5, no α-site ligand, active species	$5.67s^{-1}$	[173]
$E(A\text{-}A) \rightarrow E(Q_1)$	open	pH = 6.5, no α-site ligand, active species	$2.23s^{-1}$	[173]
$E(Q_1) \rightarrow E(A\text{-}A)$	open	pH = 6.5, no α-site ligand, inactive species	$4.03s^{-1}$	[173]
$E(A\text{-}A) \rightarrow E(Q_1)$	open	pH = 6.5, no α-site ligand, inactive species	$0.18s^{-1}$	[173]
$E(A\text{-}A) \rightarrow TS\text{-}Ser$	not specified		$10s^{-1}$	[169]
$E(A\text{-}A) + Ind \rightarrow E(A\text{-}A)^*\text{-}Ind$	not specified		$2\mu M^{-1}s^{-1}$	[169]
$E(A\text{-}A) + Ind \rightleftharpoons E(A\text{-}A)^*\text{-}Ind$	not specified	depends on NaCl concentration, [NaCl] = 0	$K = 1.5mM^{-1}$	[45]
$E(A\text{-}A) + Ind \rightleftharpoons E(A\text{-}A)^*\text{-}Ind$	not specified	depends on NaCl concentration, [NaCl] = 100mM	$K = 3.8mM^{-1}$	[45]
$E(A\text{-}A) + Ind \rightleftharpoons E(A\text{-}A)^*\text{-}Ind$	not specified		$K = 6.4 \cdot 10^4 M$	[43]
$E(A\text{-}A)\text{-}Ind \rightarrow E(Q_2)$	not specified	no α-site ligand, two possible reaction pathways	$250s^{-1}$	[168]
$E(A\text{-}A)\text{-}Ind \rightarrow E(Q_2)$	not specified	with IPP, two possible reaction pathways	$50s^{-1}$	[168]
$E(A\text{-}A)\text{-}Ind \rightarrow E(Q_2)$	closed		$365s^{-1}$	[43]
$E(Q_2) \rightarrow E(A\text{-}A)\text{-}Ind$	closed		$25s^{-1}$	[43]
$E(A\text{-}A)^*\text{-}Ind \rightarrow E(Q_3)$	closed		$270s^{-1}$	[45]
$E(Q_3) \rightarrow E(A\text{-}A)^*\text{-}Ind$	closed	depends on NaCl concentration, [NaCl] = 0	$20s^{-1}$	[45]
$E(Q_3) \rightarrow E(A\text{-}A)^*\text{-}Ind$	closed	depends on NaCl concentration, [NaCl] = 100mM	$1s^{-1}$	[45]
$E(Q_3) \rightarrow E(Aex_2)$	not specified		$50s^{-1}$	[45]
$E(Q_2) \rightarrow E(Aex_2)$	not specified	no α-site ligand, two possible reaction pathways	$14s^{-1}$, $5.0s^{-1}$	[168]

$E(Aex_2) \rightarrow E(Q_2)$	not specified	no α-site ligand, two possible reaction pathways	$7.8s^{-1}$, $1.7s^{-1}$	[168]
$E(Q_2) \rightarrow E(Aex_2)$	not specified	with IPP, two possible reaction pathways	$6s^{-1}$, $0.17s^{-1}$	[168]
$E(Aex_2) \rightarrow E(Q_2)$	not specified	with IPP, two possible reaction pathways	$2s^{-1}$, $1.9s^{-1}$	[168]
E(A-A)*-Ind $\rightarrow$ TS*-Trp	closed		$\geq 1000s^{-1}$	[169]
TS*-Trp + $\rightarrow$ E(A-A)*-Ind	not specified		$0.1s^{-1}$	[169]
$E(Aex_2) \rightarrow$ TS + Trp	not specified	no α-site ligand	$40s^{-1}$	[168]
TS + Trp $\rightarrow E(Aex_2)$	not specified	no α-site ligand	$1.5 \cdot 10^5 M^{-1}s^{-1}$	[168]
$E(Aex_2) \rightarrow$ TS + Trp	not specified	with IPP	$30s^{-1}$	[168]
TS + Trp $\rightarrow E(Aex_2)$	not specified	with IPP	$0.3 \cdot 10^5 M^{-1}s^{-1}$	[168]
TS*-Trp $\rightarrow$ TS + Trp	switching probably included		$8s^{-1}$	[169]
TS + Trp $\rightarrow$ TS*-Trp	switching probably included		$0.5\mu M^{-1}s^{-1}$	[169]

Table 2.2: β-Reaction: kinetic rate constants. The data from [44] was fitted to the following mechanism: TS + Ser $\rightleftharpoons$ TS-Ser $\rightleftharpoons E(Aex_1) \rightleftharpoons$ E(A-A). The constants from [45] were obtained fitting the simplified mechanism: Ind + E(A-A) $\rightleftharpoons$ E(A-A)-Ind $\rightleftharpoons E(Q_3) \rightleftharpoons E(Aex_2)$. The data from [43] was obtained fitting the mechanism: E(A-A) + Ind $\rightleftharpoons$ E(A-A)-Ind $\rightleftharpoons E(Q_2)$. The results from [169] were obtained using KINSIM. In [173], the addition of an α-site ligand (indole propanol phosphate) shifts the equilibrium distribution between the active and inactive forms of $E(Aex_1)$. In [168], the reaction proceeds without tunneling, because indole is used instead of IGP; nevertheless, the influence of the α-site ligand indole propanol phosphate is investigated. Unless stated otherwise, the experiments were performed under pH = 7.6 Abbreviations: TS: tryptophan synthase, IPP: indole propanol phosphate, KINSIM: kinetic simulation program [171].

2.3 Construction of the Single-Molecule Model

As discussed in the end of section 1.2, when modeling enzyme dynamics including chemical reactions, the nature of the model depends on the ratio of the time scales of conformational motions and chemical reactions. In the case of tryptophan synthase, the time scale of the slowest chemical reactions is on the order of 0.1 s. The characteristic time scale for large scale conformational motions in motor proteins is known to be on the order of milliseconds [100, 99]. To the author's knowledge, no direct measurements of the time scale of conformational motions for tryptophan synthase are available in the literature. However, it is safe to assume that the conformational motions in tryptophan synthase are

not slower than the motions in motor proteins and therefore are considerably faster than the chemical reactions. This is a rather unusual situation for a protein machine - usually the ratio is reversed. However, the sequence of reactions catalyzed by tryptohan synthase is very complex and includes diverse C-C and C-N bond formations and cleavages in different positions of the PLP-substrate complex (figure 1.8). Therefore, the catalytic center cannot be optimized for all the elementary reactions, but is a compromise in terms of overall performance resulting in unusually slow rates for some reactions. In effect, the conformational motions can be integrated out and the natural coarse-grained model for tryptophan synthase is a Markov network with discrete states given by the chemical states of both α- and β-subunits and transitions corresponding to chemical reactions within either one of the subunits.

The construction of a single-molecule Markov model requires the explicit identification of all states (a, b) of the Markov network, where a is a chemical state of the α-subunit and b is a chemical state of the β-subunit. Moreover, the transition probability rates $w_{a,a'}^{b,b'}$ from the state (a', b') to (a, b) need to be identified. The starting point is experimental data given by the set of all chemical states a and b of the α- and β-sites and all rates for transitions $a \to a'$ and $b \to b'$ measured in diverse experiments (many transitions have more than one measured rate due to measurements under different conditions). Note that the raw data for the transitions is not known for combined states (a, b), but only for individual states a and b. A set of suitable rate constants is identified for each subunit (**Step 1**). Then states that form and decay fast are adiabatically eliminated (**Step 2**). At this stage, the combined states (a, b) are introduced (**Step 3**). From experiments it is known that some combinations (a, b) of chemical states are not possible due to conformational limitations (**Step 4**). Moreover, certain transitions need to be modified due to the allosteric interactions and conservation of mass preventing the loss or spontaneous appearance of indole inside the enzyme (**Step 5**). Fixing a set of substrate and product concentrations gives the final Markov network model (**Step 6**).

Step 1

All the chemical species of both subunits are well-known and can all be found in [19]. They are shown in figure 1.8. The experimentally measured reaction rates depend on the experimental conditions and choices are made based on the following principles: Studies where several rate constants have been determined within the same experimental setup are preferred. Moreover, because allosteric interactions between the two subunits are important, only experimental results for the full $\alpha_2\beta_2$ enzyme are used. The experiments by Anderson *et al.* [169] and Lane and Kirschner [168] yield most of the data to determine the transition rates in the stochastic Markov network model. However, the experimental conditions in these investigations were still not identical: while Anderson *et al.* [169] used the enzyme from *Salmonella Typhimurium* at 37°C, the experiments [168] were performed with the *Escherichia Coli* enzyme at 25°C.

Step 2

The α-reaction The data for the α-reaction in the model is taken from the work by Anderson *et al.* [169]. For the α-site, the reaction sequence empty $\rightleftharpoons$ IGP $\rightleftharpoons$ G3P+indole

$\rightleftharpoons$ G3P $\rightleftharpoons$ empty was assumed by these authors. Later, it has been found that the cleavage of IGP at the α-site is a concerted reaction [35, 36, 29] and therefore Anderson *et al.* [169] have indeed correctly taken into account all reaction steps at the α-site. Transition rate constants based on [169] are given in figure 2.2A. To compute them, the concentration of the substrate IGP was chosen the same as in the experiments [169] and the product G3P concentration was set to zero (see table 2.4). No adiabatic elimination was performed for the α-reaction sequence. In ref. [169], the reaction sequence at the β-site has been treated only in a simplified way: all intermediates except for the aminoacrylate A-A were merged into single enzyme-substrate and enzyme-product complexes. Therefore, while the data by Anderson *et al.* [169] was sufficient to model the α-reactions, other results were used for the β-reactions.

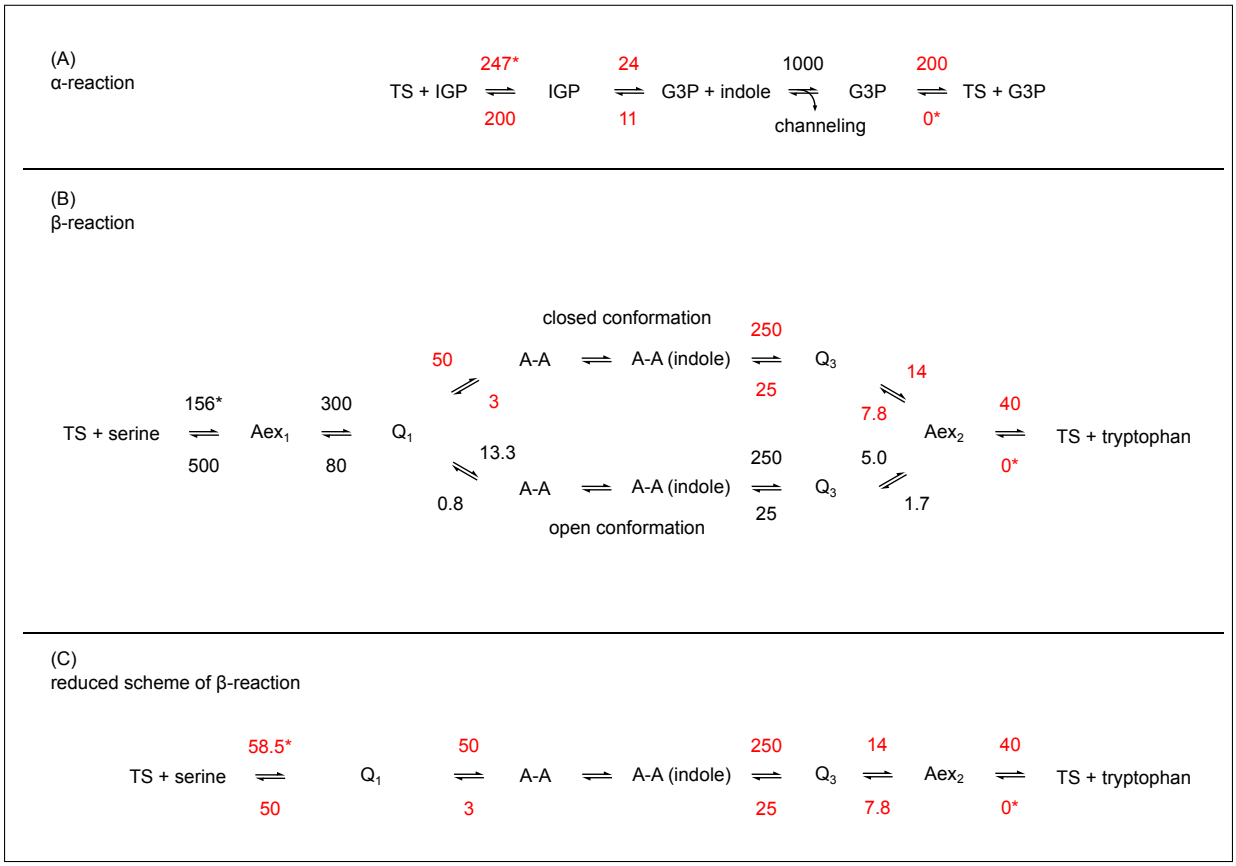


Figure 2.2: Reactions at α- and β-sites and their rate constants [s^{-1}]. The values in red were used in the Markov network model. The constants marked with an asterisk are the first-order rate constants obtained from the second-order rate constants by fixing substrate and product concentrations from table 2.4.

The β-reaction In comparison to the α-reaction, the reaction mechanism at the β-site is considerably more complex. The rate constants obtained by Lane and Kirschner [168] provide the basis for the model constructed here. Lane and Kirschner have investigated the kinetics of the β-reaction in the $\alpha_2\beta_2$ enzyme and resolved all known intermediates except only for two geminal diamines GD_1 and GD_2 that were too fast to be observed. Based on their measurements, branching of the β-reaction pathway in the part corresponding to the

α-site	empty	IGP	indole+G3P	G3P		
Variable a	1	2	3	4		
β-site	empty	Q_1	A-A	indole+A-A	Q_3	Aex_2
Variable b	1	2	3	4	5	6

Table 2.3: Chemical states at the α- and β-sites after adiabatic elimination and enumeration by variables a and b.

reaction sequence A-A $\rightleftharpoons$ A-A(indole) $\rightleftharpoons$ Q_3 has been proposed (figure 2.2B). However, several years later and with improved experimental techniques, Woehl and Dunn [44, 45] have come to the conclusion that the branch corresponding to the closed β-subunit plays the dominant role. In the present study, following [44, 45] the other branch of the pathway is discarded (figure 2.2C). The intermediate Aex_1 is short-lived, with a decay rate of 800 s^{-1}. In the reduced model (figure 2.2C), it has been adiabatically eliminated yielding the apparent rate constants $k_+ = 156 \cdot 300/800\ s^{-1} = 58.5\ s^{-1}$ for the transition TS + serine $\rightarrow$ Q_1 and $k_- = 80 \cdot 500/800\ s^{-1} = 50\ s^{-1}$ for the respective backward transition. Note that Lane and Kirschner [168] have performed experiments under two different pH-values of 6.5 and 7.6. To match the experiments [169], the data determined at pH = 7.6 was used. Furthermore, rate constants obtained in absence of α-site ligands were chosen.

The reaction rate constant for the transition $Q_3 \rightarrow$ A-A(indole) could not be determined by Lane and Kirschner. However, it was found by Leja *et al.* [43]. The rate constants for binding and release of substrates and products have been determined in [177]. For each catalytic site, we thus obtain a full set of reaction rate constants shown in red in figure 2.2. Note that the rate constant for reverse indole channeling has not been determined experimentally to the author's knowledge. In the experimental literature cited in this thesis, it is generally assumed to be irreversible.

The chemical states to be included in the model after adiabatic elimination are given in table 2.3. For notational convenience, numerical variables a and b are introduced in this table.

Step 3

The chemical state of a single molecule of tryptophan synthase is given by its states a and b at both the α- and β-site. The complete *unrestricted* state space is thus given as the space of *combined* states $\{(a,b)|a = 1,2,3,4; b = 1,2,...,6\}$. It is shown in figure 2.3. Transitions correspond to chemical reaction at either the α- or the β-site. Note that for chemical reactions, simultaneous transitions of both sites need not be included in the network, because such transition probability rates within a time interval dt are on the order of dt^2. Simultaneous transitions are only important if the states of both sites change at the same time due to indole channeling.

Figure 2.3 corresponds to two noninteracting subunits. However, as described in the introductory section 1.1, the reaction at the catalytic sites of both subunits are tightly coupled.

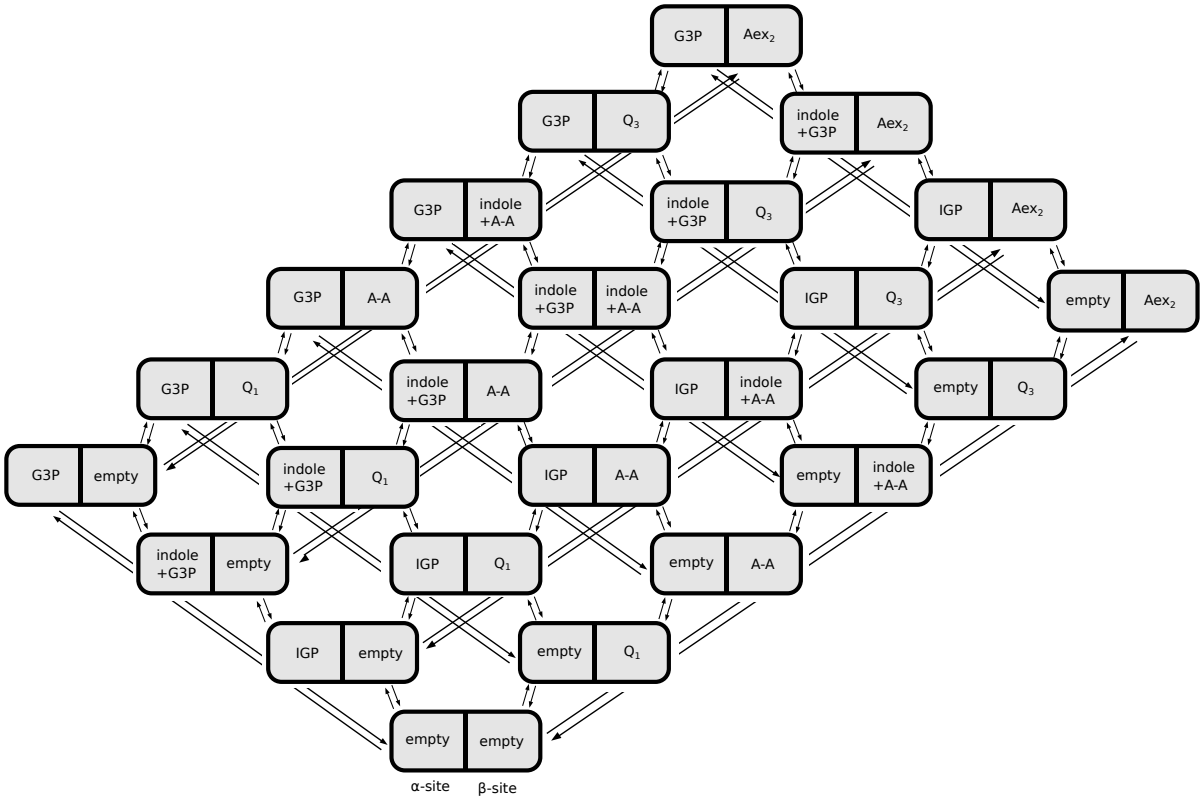


Figure 2.3: The unrestricted state space of combined states $\{(a, b), a = 1, ..., 4; b = 1, ..., 6\}$ and all possible transitions.

Step 4

Each catalytic site has a gating mechanism that prevents the exchange of matter with the environment. In the α-subunit, the loops αL2 and αL6 can fold over the entrance to the catalytic site [32, 31]. In the β-subunit, βAsp305 rotates and establishes a hydrogen bonding network with the surrounding residues to close the gate to the catalytic site [28, 38]. The enzyme adopts only two conformational states with either both gates open (open conformation) or closed (closed conformation) [44, 31, 19]. The preferred conformation of the enzyme is determined by the chemical states at both catalytic sites. The assignment of conformations to chemical states based on crystallographic experiments has been discussed in Refs. [44, 31]. The β-states empty, Q_1 and Aex_2 correspond to open conformation, while the β-states A-A and Q_3 have the conformation with both gates closed. The α-state empty is only found in the open conformation, the state indole+G3P is present only in the closed conformation and the states IGP and G3P can adopt both open and closed conformations. Therefore, the combinations (empty,A-A), (empty,A-A(indole)), (empty,Q_3), (IGP,A-A(indole)) (indole+G3P,empty), (indole+G3P,Q_1), (indole+G3P,A-A(indole)) and (indole+G3P,Aex_2) do not occur due to the incompatibility of conformations of the two subunits. The resulting *reduced* state space is shown in figure 2.4.

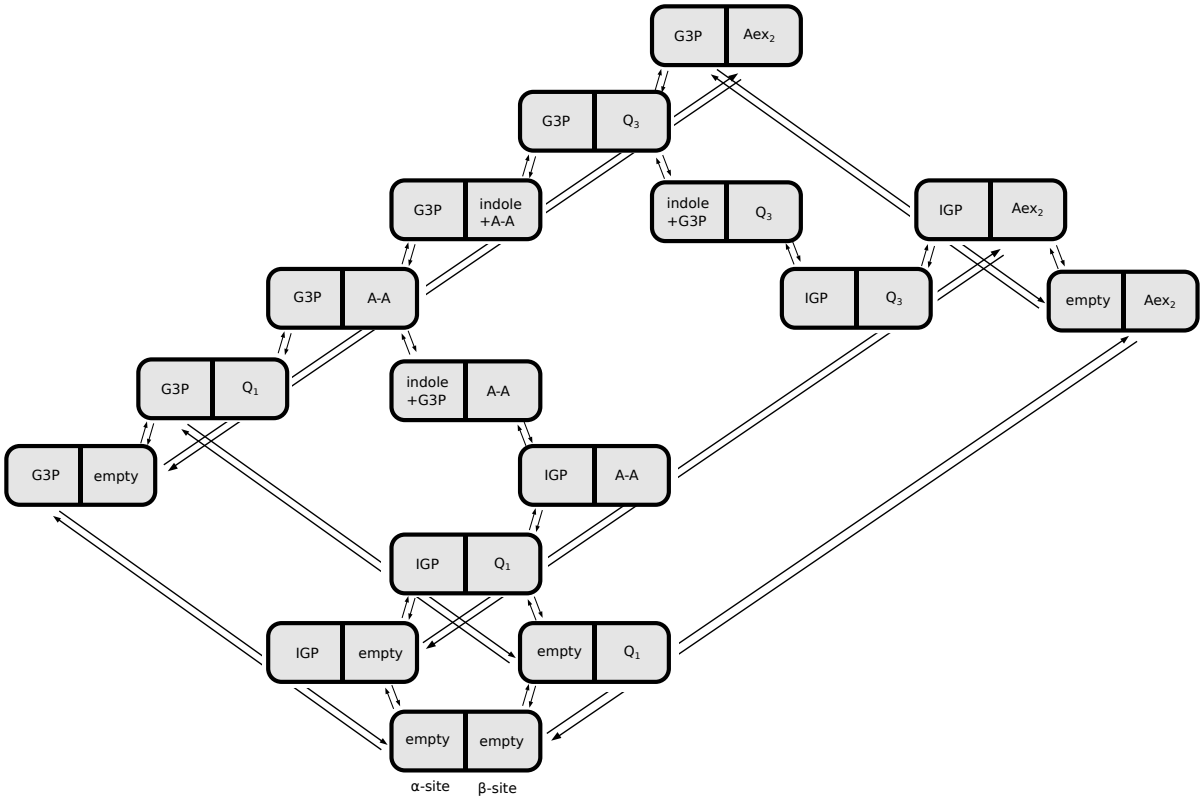


Figure 2.4: The reduced state space of combined states (a, b) with the combinations of states (empty,A-A), (empty,A-A(indole)), (empty,Q_3), (IGP,A-A(indole)) (indole+G3P,empty), (indole+G3P,Q_1), (indole+G3P,A-A(indole)) and (indole+G3P,Aex_2) excluded due to incompatibility of the α- and β conformational states.

Step 5

The reactions at α- and β-sites are coupled through indole channeling and also through allosteric interactions. The rate of indole channeling (1000 s^{-1}) is taken from ref. [169] (figure 2.2). Allosteric interactions between the sites lead to reaction rate enhancements and control the gates for arrival of substrates and release of products at both catalytic sites. The reaction rate enhancements have been studied in kinetic experiments [29, 42], where actual ligands were replaced by structurally similar, but unreactive analogues. The presence of IGP at the α-site increases the rate of formation of the aminoacrylate A-A at the β-site by a factor of 9.7 [29] When A-A is present at the β-site, this activates the cleavage of IGP at the α-site by a factor of 27.7 [42].

Because indole channeling is fast, indole release at the α-site (indole+G3P $\rightarrow$ G3P) and indole uptake at the β-site (A-A $\rightarrow$ A-A (indole)) take place only simultaneously. Moreover, the closed states (IGP,A-A), (G3P,A-A), (IGP,Q_3) and (G3P,Q_3) cannot release IGP or G3P from the α-site. The allosteric activations are modeled by multiplying the transition rates for Q_1 $\rightarrow$ A-A and IGP $\rightarrow$ indole+G3P by 9.7 and 27.7 in the transitions (IGP,Q_1) $\rightarrow$ (IGP,A-A) and (IGP,A-A) $\rightarrow$ (indole+G3P,A-A) on the network. This yields a modification of the possible transitions on the network of combined states resulting in

the reaction network shown in figure 2.5. The states are colored according to the preferred conformation.

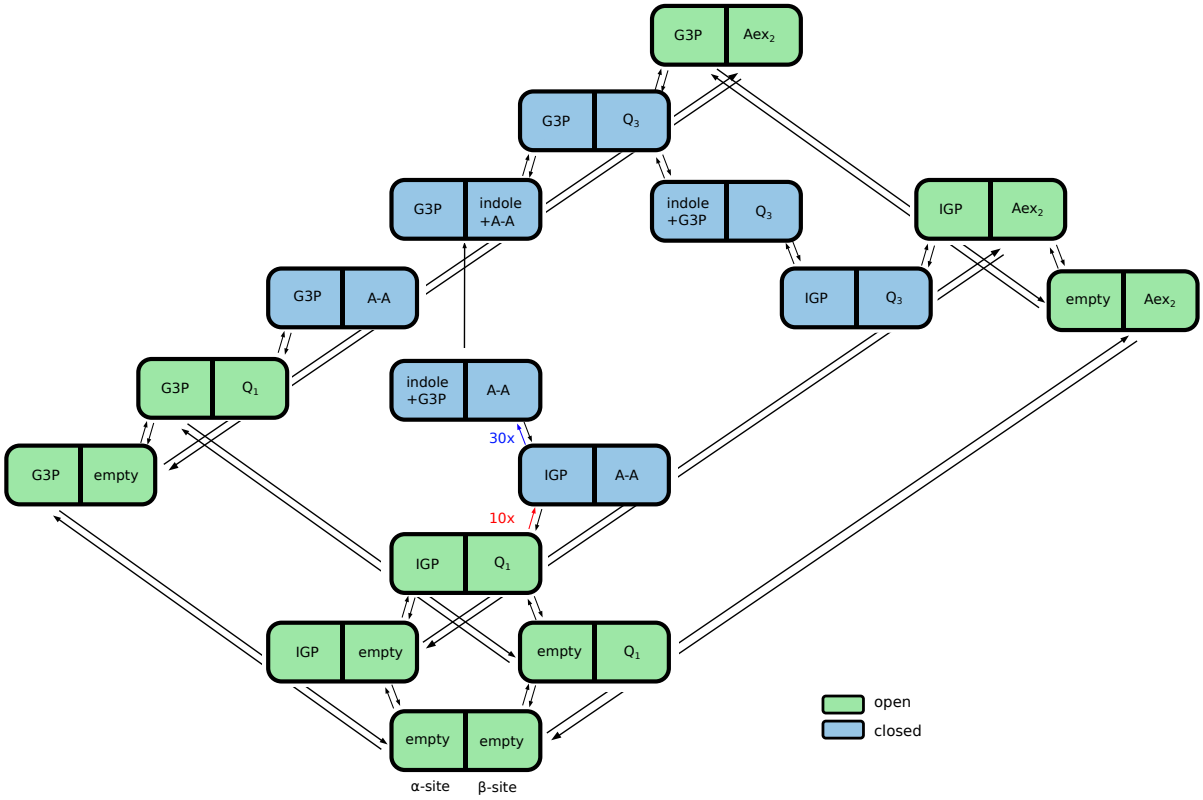


Figure 2.5: Reaction network for on the reduced states space of combined states taking into account the allosteric interactions (red and blue arrows), the impossibility of substrate binding and release in the closed conformation and indole channeling as a simultaneous transition of both sites. The closed states are colored blue and the open states are shown in green.

Step 6

In a typical experimental situation product concentrations remain vanishingly small and thus the product binding rates are assumed to be zero in this chapter. Moreover, the substrate concentrations are taken from the same sources as the majority of the rate constants [169, 168]. The numerical values are given in table 2.4.

Reaction	Rate k	Concentration c	Rate $\tilde{k}$	Source
TS + Ser $\rightarrow$ E(Q_1)	$7.5 \cdot 10^{-2} \mu M^{-1} s^{-1}$	c(Ser)=2.08 mM	156 s^{-1}	[168]
TS + Trp $\rightarrow$ E(Aex_2)	$0.15 \mu M^{-1} s^{-1}$	c(Trp)=0	0	[168]
IGP + TS $\rightarrow$ IGP-TS	$10 \mu M^{-1} s^{-1}$	c(IPG)=24.7μM	247 s^{-1}	[169, 177]
G3P + TS $\rightarrow$ G3P-TS	$0.2 \mu M^{-1} s^{-1}$	c(G3P)=0	0	[169]

Table 2.4: Measured second-order rate constants k and the respective first-order constants $\tilde{k}$ for the chosen concentrations c. The first-order constants were computed as $\tilde{k} = k \cdot c$.

This completes the construction of the Markov network model of tryptophan synthase.

2.4 Kinetic Markov Network Model

The kinetic model model from figure 2.5 is redrawn in figure 2.6A with emphasis on the role of the different states in the catalytic cycle. The large green boxes correspond to substrate and product binding and release. The states involved in these reactions have open conformations. The main catalytic functions are carried out in closed states. Hereby, the mutual allosteric activations, the indole channeling from the α- to the β-site and reaction of indole at the β-site form a catalytic chain (large blue box). Note the presence of the "futile" states (indole+G3P,Q_3) or (G3P,A-A) (orange boxes in figure 2.6A). In these states, the enzyme cannot catalyze any fertile reactions, because it either contains two indole equivalents (state (indole+G3P,Q_3)) or no indole equivalents (state (G3P,A-A)). Thus, to proceed further with fertile catalytic reactions, the enzyme has to return to an open conformation to release the product bound at one catalytic site and bind new substrate. "Futile" states do not contribute to the catalytic reaction, but lead to an increase of the turnover time.

Alternatively, the same stochastic model can be formulated in terms of the two interacting Markov chains for the α- and β-sites (figure 2.6B). The reactions modify the states of one subunit, but they can be enhanced or inhibited (blocked) depending on the state of the other subunit. Additionally, there is one reaction (i.e., indole channeling) which simultaneously changes the states of both subunits.

Within its catalytic cycle, the tryptophan synthase molecule undergoes a sequence of reaction events each associated with a change of its chemical state. This sequence can be considered as a random walk over the set of states (a, b). The set of states together with the possible transitions between them define a Markov network. For the combined states (a, b), time dependent probabilities $p(a, b; t)$ can be introduced. They satisfy the master equation

$$\frac{d}{dt}p(a,b;t) = \sum_{a'=1}^{4}\sum_{b'=1}^{6}[w_{a,a'}^{b,b'}p(a',b';t) - w_{a',a}^{b',b}p(a,b;t)] \tag{2.4.1}$$

introduced in section 1.3.

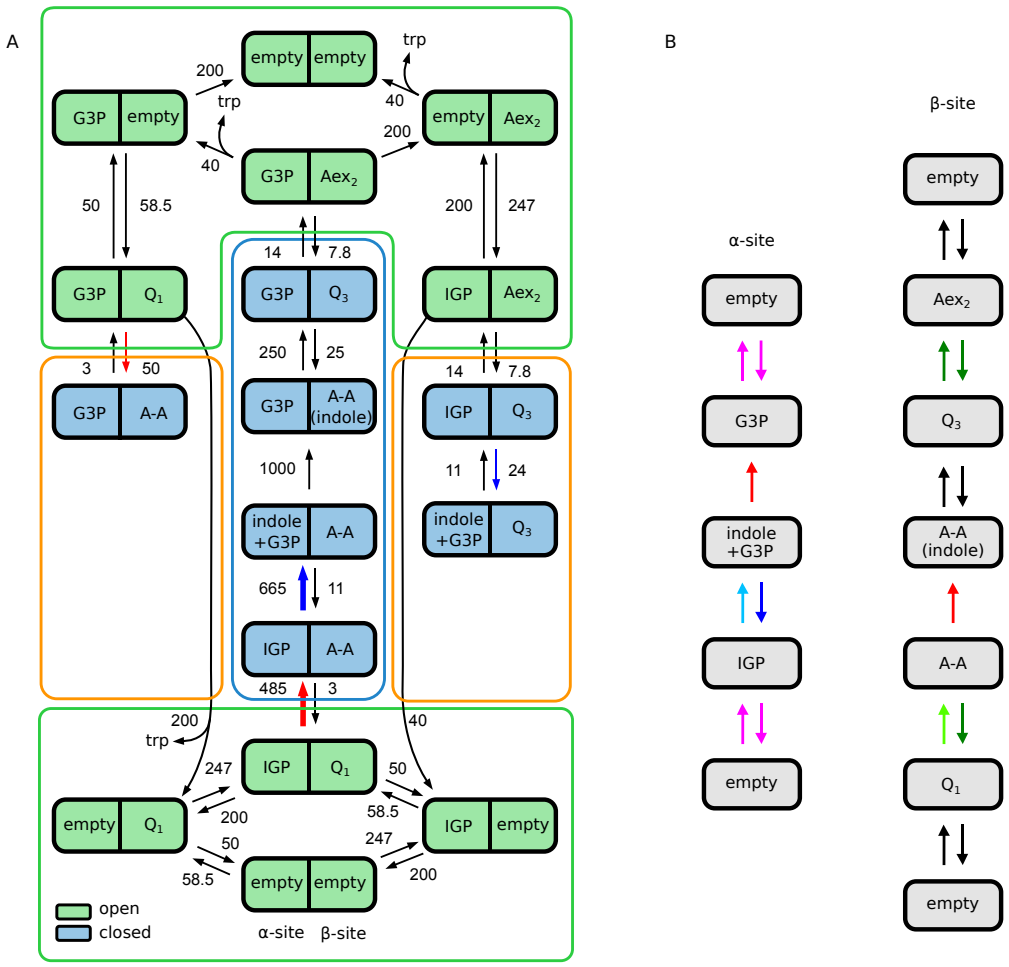


Figure 2.6: The Single-molecule stochastic model of tryptophan synthase. **A** Markov network with numerical values of transition rates [s^{-1}]. Green and blue colors correspond to open and closed conformations. **B** Equivalent representation as two interacting Markov chains. Magenta: transitions blocked in the states A-A and Q_3 of the β-site. Green (light and dark): blocked in the state empty of the α-site. Light green: enhanced by a factor of 9.7 in the state IGP of the α-site. Blue (light and dark): blocked in the states empty, Q_1, Aex_2 of the β-site. Light blue: enhanced by a factor of 27.7 in the state A-A of the β-site. Red: Channeling instantaneously changes the states of both sites.

2.5 Simulation Results

In stochastic numerical simulations, the chemical reaction course inside a single tryptophan synthase enzyme is reproduced using the Gillespie algorithm [181]. Starting from the state (empty,empty), the enzyme performs a random walk on the Markov network shown in figure 2.6A. This walk represents a series of transitions whose probability rates are all known. The cycle ends when both products are released and the enzyme returns to its initial state. An example of a 2.13 s time series is shown in figure 2.7A. In the simulations, numerical data for one million turnover cycles has been collected and analyzed.

Figure 2.7B shows the distribution of overall turnover times for tryptophan synthase.

The mean turnover time is $\mu = 0.15$ s. However, it has a thin long tail of cycle durations on the order of several seconds. This tail is a result of stochastic fluctuations that drive the two catalytic sites out of phase and lead to prolonged retention in the "futile" states (indole+G3P,Q_3) and (G3P,A-A). It has been checked that, if the transitions to all "futile" states (orange boxes in figure 2.6A) are blocked, the tail disappears.

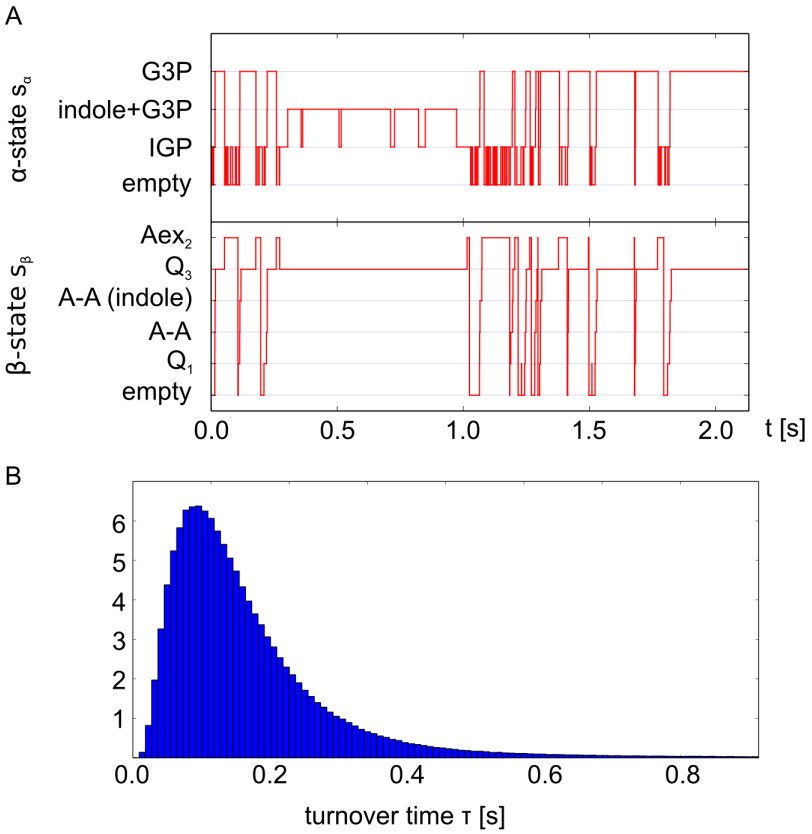


Figure 2.7: Simulation data. **A** Example of a short time series from a of 2.13 s duration. Horizontal red lines indicate the enzyme being in the respective state and vertical red lines indicate transitions between the states. **B** Normalized histogram of turnover times using the data for 10^6 cycles. **C** Histogram shown in **B** plotted with a logarithmic scale reveals a tail of long cycle durations.

Using the simulation data, joint probabilities $p(a, b)$ to find the enzyme in different compatible combinations of internal states (a, b) were determined. Joint occupation probabilities $p(a, b)$ for different states a and b were always obtained from stochastic simulations of 10^6 turnover cycles. These probabilities are displayed in figure 2.8; their numerical values are given in table B.1. Once both substrates have arrived, the enzyme quickly proceeds to indole formation and channeling. After that, it stays however for a long time in the state (G3P,Q_3). The probabilities $p(a)$ and $p(b)$ to find the enzyme in the states a and b irrespectively of the states at the other subsite can be obtained by summing $p(a, b)$ over all states of the other subsite (see table B.2 for their numerical values).

If the α- and β-subunits of the enzyme were independent chemical species, the joint probability distribution $p(a, b)$ would have been given by a product of the probabilities

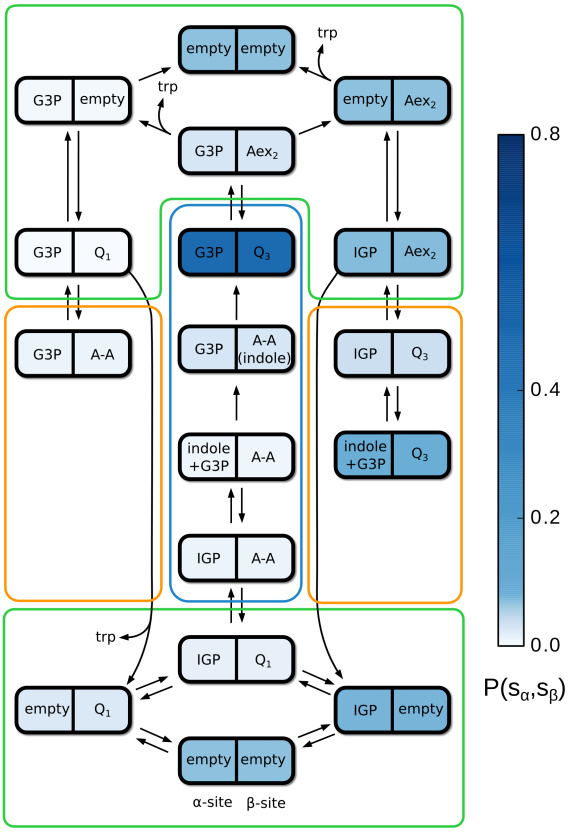


Figure 2.8: Joint probabilities $p(a, b)$. Numerical values are given in table B.1.

$p(a)$ and $p(b)$. Hence, intramolecular correlations between the internal states of the two subunits can be characterized by the difference between the joint probability $p(a, b)$ and the product $p(a)p(b)$. It is convenient to normalize this difference in such a way that the resulting correlation coefficient $c(a, b)$ can vary only from -1 to +1, taking the extreme values in the case of completely correlated or anti-correlated states. For any two chosen states a and b of α- and β-subunits, the random binary variables $X(a)$ and $X(b)$ are defined such that they take values 1 if the respective subunit is in the chosen state and zero otherwise. The elements of the correlation matrix $c(a, b)$ are defined as Pearson correlation coefficients of the random variables $X(a)$ and $X(b)$, i.e. as

$$c(a,b) = \frac{\langle X(a)X(b)\rangle - \langle X(a)\rangle\langle X(b)\rangle}{\sqrt{\langle X(a)^2\rangle - \langle X(a)\rangle^2}\sqrt{\langle X(b)^2\rangle - \langle X(b)\rangle^2}}, \tag{2.5.1}$$

where $\langle . \rangle$ denotes the ensemble averaging. Thus defined, the correlation coefficients take the maximal value of 1 if $X(a) = X(b)$ and the minimal value of -1 if $X(a) = -X(b)$. They are expressed in terms of the occupation probabilities as

$$c(a,b) = \frac{p(a,b) - p(a)p(b)}{\sqrt{p(a) - p(a)^2}\sqrt{p(b) - p(b)^2}}, \tag{2.5.2}$$

where $p(a, b)$ is the joint probability to find the two enzyme subunits in the respective states and $p(a) = \sum_b p(a, b)$, $p(b) = \sum_a p(a, b)$.

The computed intramolecular correlation matrix is displayed in figure 2.9. The strongest correlation (0.61) is found between the states G3P and Q_3. Indeed, both subunits arrive almost simultaneously to this state, due to reciprocal strong allosteric activation along the main catalytic pathway. The substantial correlations or anti-correlations involving empty states of both subunits are due to allosteric opening or closing of the gates.

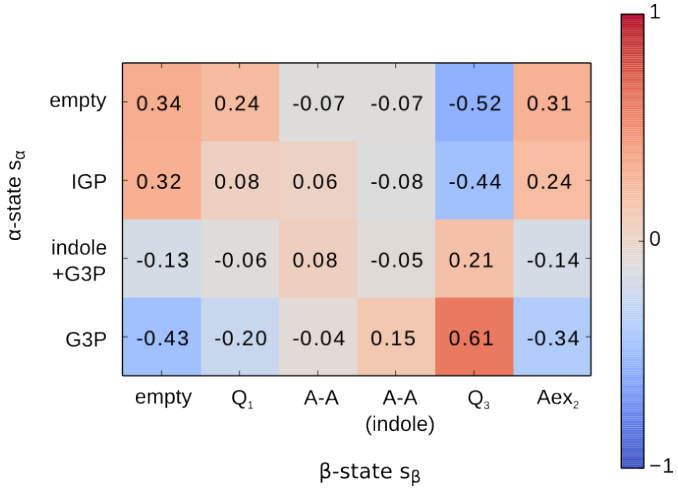


Figure 2.9: Intramolecular correlations $c(a, b)$ between different internal states of the two subunits. In addition to color coding (see the bar), numerical values of the Pearson correlation coefficients are also given.

For any α- or β-state s, the first-passage time $t(s)$ can be defined as the time the enzyme needs to reach this state s when starting from the initial state where both catalytic sites are unoccupied. The mean-square-root time dispersion $\sigma(a, b) = \langle (t(a) - t(b))^2 \rangle^{1/2}$ of first-passage times for any two states a and b characterizes the degree of temporal synchronization between these two states. The binding of substrates can take place in an arbitrary order and is not controlled by allosteric interactions. The simulations yield $\sigma(\text{IGP},Q_1) = 22$ ms for the temporal correlation between binding of two substrates to their respective α- and β-sites. In comparison σ(indole+G3P,A-A) = 2.4 ms, and thus the states before indole channeling are reached almost simultaneously at both catalytic sites. This clearly demonstrates the buildup of synchronization in tryptophan synthase.

The stochastic model can be used not only to reproduce the actual operation of tryptophan synthase, but also to perform *in silico* studies of its operation mechanism and of the role of allosteric regulation in its function. As shown in figure 2.6, there are two reactions steps which are allosterically activated, i.e. the transitions of Q_1 to A-A and of IGP to indole+G3P. How does the action of the enzyme at the single-molecule level change if both allosteric regulations are switched off or both permanently activated?

To answer this question, simulations in the absence or permanent presence of both activations have been performed. They show that the mean turnover time of the native enzyme ($\mu = 0.15$ s) is about two times shorter than that of the hypothetical enzyme with absent ($\mu = 0.26$ s) and more than three times shorter than that of the hypothetical enzyme with permanently present ($\mu = 0.52$ s) activations.

While the increase of the turnover times in absence of activations is well expected, since some transitions in the main catalytic pathway become slower, their increase under permanent activations needs further analysis. Figure 2.10 shows occupation probabilities of different enzyme states in such two cases. Comparing figure 2.10A with figure 2.8A, it can be noticed that, in absence of activations, the enzyme spends more time in the states (IGP,Q_1) and (IGP,A-A), the transitions from which are slowed down. When both activations are permanently present (figure 2.10B), the occupations probabilities of these states become close to those for the native enzyme (figure 2.8A). However, the enzyme now spends much time in the futile state (indole+G3P,Q_3). This explains a decrease in the catalytic efficiency when both allosteric activations are permanently present.

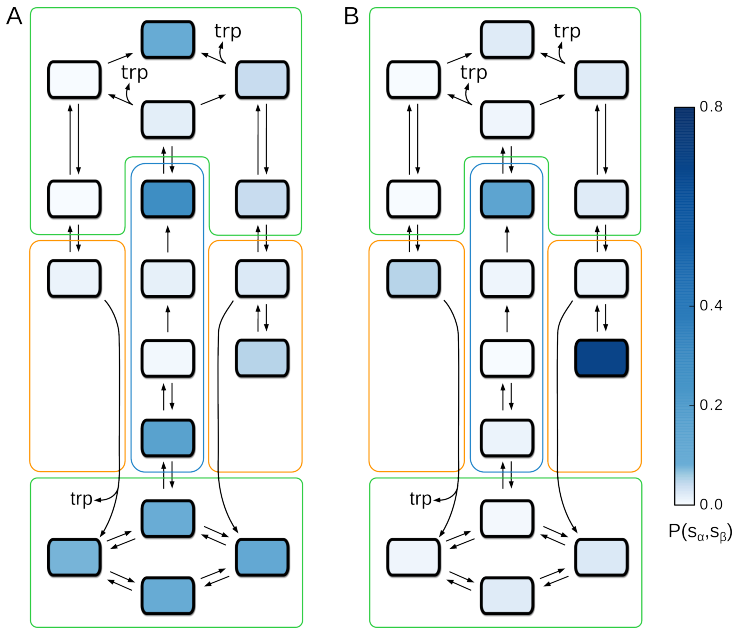


Figure 2.10: Joint probabilities $p(a, b)$ of the states of α- and β-subunits. **A** without activations and **B** with permanent activations.

The turnover time distribution of the enzyme with permanent activations closely resembles the distribution of the native enzyme for small turnover times. In these cases, the enzymes do not enter "futile" states. However, the histogram of the enzyme with permanent activations has a long tail of cycle durations ranging up to 14 s (Figure 2.11). Once a "futile" state is reached, this enzyme can dwell there for a long time thus decreasing the catalytic efficiency. Hence, the controlled activation of reaction events along the main catalytic pathway allows tryptophan synthase to raise its turnover efficiency as compared to permanent acceleration of the respective reaction events.

2.6 Discussion

While presence of strong correlations and synchronization of chemical reaction events at two catalytic subunits in tryptophan synthase has been suggested as the distinguishing feature of this chemical nanomachine (see reviews [31, 19]), such effects could not

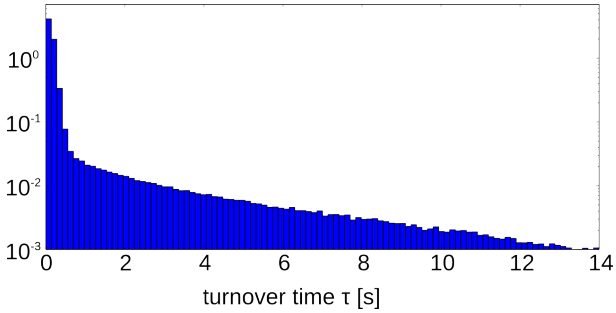


Figure 2.11: Normalized histogram of turnover times for the hypothetical enzyme with permanent activations. Data for 10^6 turnover cycles. The log-scale representation is chosen.

be accounted for in previous kinetic studies [169] where the two subunits were treated as separate chemical species. In contrast, the first stochastic single-molecule model was constructed and investigated that allows a detailed exploration of intramolecular synchronization phenomena.

Because tryptophan synthase has been broadly investigated in the past, providing indeed a "mine for enzymologists" [182], all model parameters could be extracted from the available experimental data. Through numerical simulations of the developed stochastic model, the statistics of turnover cycles in this enzyme could be determined. The predicted mean turnover time under the saturation concentrations was found to be equal to 0.15 s which is comparable with the values of 0.20 [172] and 0.30 s [42] reported under different experimental conditions. It was found that the distribution of turnover times possesses a long tail and, with significant probability, turnover cycles with the duration of a few seconds should also be observed. They are explained by dwelling of the enzyme in the futile states where the catalytic conversion becomes blocked. The dependence of the turnover rate on substrate concentrations is discussed in [21], supplementary information.

The model yields direct theoretical evidence for intramolecular synchronization phenomena. It is found that correlations between instantaneous chemical states of the two catalytic subunits can be as high as 0.61, while the absence of correlations corresponds to the zero value and the maximal possible correlation level is one. It could also be seen how temporal correlations become enhanced along the main catalytic pathway in the enzyme molecule, with the mean-square-root time dispersion falling from about 22 ms for the arrival of substrates to only about 2 ms for the arrival of intermediate products for the final catalytic conversion event.

By using the Markov model, the aspects of catalytic efficiency and allosteric regulation in tryptophan synthase could furthermore be explored. While intramolecular channeling of indole is already strongly contributing towards the efficiency by preventing its loss in a biological cell and minimizing the time needed for the transfer of this intermediate from one catalytic center to another, complex allosteric regulation contributes to further efficiency gains. Particularly, this allows to avoid, to a large extent, dwelling in the futile states which correspond to the non-productive side branches of the intramolecular catalytic pathway and thus accelerates the overall catalytic conversion.

Despite the fact that extensive kinetic measurements and X-ray diffraction observations have been performed, tryptophan synthase has not been so far investigated in experiments with single molecules, by employing, e.g., fluorescence correlation spectroscopy [183] or FRET [184, 185] methods. Hopefully, the results of this study bring the attention to very interesting possible experiments with this enzyme, where intramolecular synchronization and the effects of strong correlations could be directly demonstrated at the single-molecule level.

As mentioned in the introduction, tryptophan synthase represents a characteristic example of a channeling enzyme and, generally, can be viewed as an analog of multienzyme complexes that play an important role in biological cells. Beyond the case of this specific enzyme, the study provides a theoretical framework for single-molecule kinetic modeling of such chemical nanofactories where entire complex catalytic pathways are efficiently implemented within one molecular nanoscale aggregate or a single oligomeric enzyme.

Chapter 3

Stochastic Thermodynamics of Tryptophan Synthase

In the previous chapter, the Markov network model has been constructed based purely on experimentally determined kinetic data. In particular, the reverse rate of indole channeling is not available experimentally and therefore was not included into the model. Moreover, the product concentrations have been set to zero to conform to experimental conditions in [168, 179, 169, 45, 178]. In this chapter, the network is analyzed using the theory of stochastic thermodynamics under physiological conditions, where the enzyme operates *far from equilibrium*. Irreversible transitions are not admissible, because they would lead to divergent values for energy differences and entropy production. Therefore, the Schnakenberg theory of cycles and fluxes together with experimental thermodynamic data is used to calculate the rate of reverse indole channeling (section 3.2) and thus to obtain a modified Markov network model with all reversible internal transitions. Moreover, physiological substrate and product concentrations are used. The results presented in this and the following chapter have been published in [25].

3.1 Preliminaries

All calculations in this and the next chapter are performed for a slightly modified version of the Markov network model constructed in the previous chapter. In the kinetic model shown in figure 2.6, the transition (indole+G3P,A-A) $\rightarrow$ (G3P,(indole)A-A) corresponding to indole channeling is irreversible in agreement with experimental observations [168, 169, 177]. This agrees with calorimetric measurements showing that tryptophan synthesis is exergonic [186]. There is a large difference in standard Gibbs free energies $\Delta_r G^0_m = -50.7$ kJ·mol^{-1} between products (G3P, tryptophan and water) and substrates (IGP and serine) in this reaction, corresponding to an energy difference of about 20 k_BT between the substrates and products. For the reverse reaction to occur, the enzyme would have to extract 20 k_BT from thermal fluctuations of its environment. This is highly improbable and therefore the reverse reaction is not observed for tryptophan synthase. The largest Gibbs free energy gap is found for the step of indole channeling (5.4 k_BT) and is calculated in the next section.

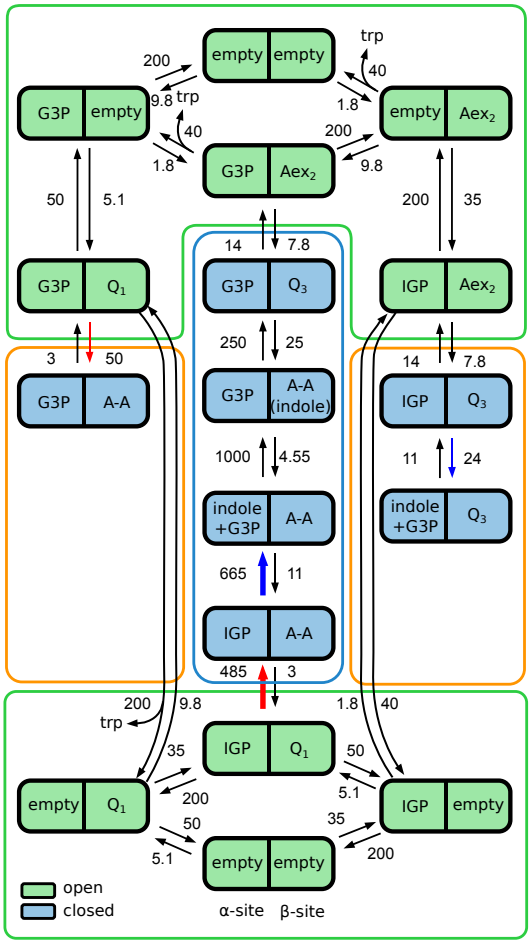


Figure 3.1: The fully reversible kinetic Markov network of tryptophan synthase with numerical values of all transition rates in units of s^{-1}.

However, from a thermodynamical point of view, irreversible reactions lead to divergent values for free energy differences, entropy production and entropy flow according to equations 1.3.16, 1.3.9 and 1.3.10

$$\frac{w_{x,x'}}{w_{x',x}} = \exp\left(\frac{F(x') - F(x) + F_{x,x'}}{k_B T}\right),$$
$$\sigma_{x,x'} = J_{x,x'} \ln \frac{w_{x,x'} p(x';t)}{w_{x',x} p(x;t)},$$
$$\text{and } h_{x,x'} = J_{x,x'} \ln \frac{w_{x,x'}}{w_{x',x}}.$$

Therefore, the network constructed previously (figure 2.6) is modified by including the transition (G3P,(indole)A-A) → (indole+G3P,A-A). Moreover, nonzero product concentrations are used. This makes the transitions corresponding to product release reversible as well. Whereas the model in the previous chapter was adjusted to a commonly used experimental setup, physiological concentrations are used here and in the next chapter. The substrate and product concentrations were determined by Bennett *et al.* [187]. The respective binding rate constants are given in table 3.1. The fully reversible Markov net-

Reaction	Binding rate constant k	Concentration c	w	Ref.
β-empty + Ser $\rightarrow$ Q_1	$7.5 \cdot 10^{-2}$ $\mu M^{-1}s^{-1}$	c(Ser)=68 μM	5.1 s^{-1}	[168]
β-empty + Trp $\rightarrow$ Aex_2	0.15 $\mu M^{-1}s^{-1}$	c(Trp)=12 μM	1.8 s^{-1}	[168]
α-empty + IGP $\rightarrow$ α-IGP	10 $\mu M^{-1}s^{-1}$	c(IGP)=3.5μM	35 s^{-1}	[169, 177]
α-empty + G3P $\rightarrow$ α-G3P	0.2 $\mu M^{-1}s^{-1}$	c(G3P)=49 μM	9.8 s^{-1}	[169]

Table 3.1: Ligand binding rate constants k, ligand concentrations c and the respective transition rates $w = kc$ under physiological conditions. The concentrations were measured by Bennett *et al.* [187].

work model with all rate constants is shown in figure 3.1.

With the same notation as in the previous chapter, the dynamics on the network obeys the master equation

$$\frac{d}{dt}p(a,b;t) = \sum_{a'=1}^{4}\sum_{b'=1}^{6}[w_{a,a'}^{b,b'}p(a',b';t) - w_{a',a}^{b',b}p(a,b;t)] \tag{3.1.1}$$

where $p(a,b;t)$ is the probability to find the enzyme in the state (a,b) at time t and $w_{a,a'}^{b,b'}$ denotes the transition probability rate from a state (a',b') to the state (a,b). As was discussed in the previous chapter (see, e.g., figure 2.6), only the transition representing indole channeling involves simultaneous changes of the states of both α- and β-sites. All other transitions change the state of only one subunit although the rates of such transitions can be controlled by the state of the other subunit. Therefore, the Markov network of tryptophan synthase has a special structure. It is almost bipartite (see [22, 23, 24]) and the transition matrix elements can be written as

$$w_{a,a'}^{b,b'} = \begin{cases} w_a^{b,b'} & \text{if } a = a' \\ w_{a,a'}^{b} & \text{if } b = b' \\ w_{4,3}^{4,3} & \text{if } (a',b') = (3,3) \text{ and } (a,b) = (4,4) \\ w_{3,4}^{3,4} & \text{if } (a',b') = (4,4) \text{ and } (a,b) = (3,3) \\ 0 & \text{else.} \end{cases} \tag{3.1.2}$$

The indole channeling couples the two subunits and perturbs the complete bipartite structure of the Markov network. Taking into account the special form (3.1.2) of the transition matrix, the master equation can also be written as

$$\frac{d}{dt}p(a,b;t) = \sum_{b'} J_a^{b,b'} + \sum_{a'} J_{a,a'}^{b} + [\delta_{(a,b)}^{(3,3)} - \delta_{(a,b)}^{(4,4)}]J^{channel}, \tag{3.1.3}$$

where $\delta_i^j = 1$, if $i = j$ and $\delta_i^j = 0$ otherwise. The fluxes corresponding to transitions inside the β-subunit are $J_a^{b,b'} = w_a^{b,b'}p(a,b';t) - w_a^{b',b}p(a,b;t)$ and the fluxes $J_{a,a'}^{b}$ for the transitions within the α-subunit are defined similarly. The flux corresponding to channeling is $J^{channel} = w_{4,3}^{4,3}p(3,3;t) - w_{3,4}^{3,4}p(4,4;t)$. The transition rate constant $w_{3,4}^{3,4}$ is now determined.

3.2 Reverse Rate of Indole Channeling

The general formalism used to determine the constant $w_{3,4}^{3,4}$ for the reverse of indole channeling has been introduced in section 1.3 and appendix A. Using Schnakenberg's theory of cycle fluxes and forces, it is possible to link the transition rate constants of a cycle to its thermodynamic force. In the case of chemical reaction networks under isothermal conditions, the only forces present are gradients of the chemical potential. In particular, the cycle fluxes in tryptophan synthase are driven by the chemical potential gradient between the products (tryptophan and G3P) and the substrates (serine and IGP). The kinetic Markov model constructed in the previous chapter has only one cycle that is driven by the chemical potential gradient and its force is precisely this gradient.

The main equations used are 1.3.15 and 1.3.17. As in section 1.3, the condition of detailed balance is that at thermal equilibrium the net probability flux between any two states is absent. For the considered network it implies that the ratio of the rates $w_{a',a}^{b',b}$ and $w_{a,a'}^{b,b'}$ for forward and backward transitions between any two states (a,b) and (a',b') satisfies the equation

$$\frac{w_{a,a'}^{b,b'}}{w_{a',a}^{b',b}} = \exp\left(\frac{G(a',b') - G(a,b)}{k_B T}\right) \tag{3.2.1}$$

where $G(a,b)$ and $G(a',b')$ are Gibbs energies of the respective states in the network at equilibrium, T is the temperature, and k_B is the Boltzmann constant. Note that in this chapter the Gibbs energies are used, whereas in section 1.3 the free energies were considered. For enzyme reactions in solution, there is neither a change of volume nor of pressure and the difference of Gibbs energies determined by equation 3.2.1 is equal to the difference of free energies given by equation 1.3.16. The reason to use the Gibbs energies here is simply that it is the natural thermodynamic state variable when considering chemical reactions. All calculations and arguments given here would work in the same way with free energies.

For transitions between the states $(a,b) \to (a',b')$ that do not involve binding or release of ligands, the rates $w_{a,a'}^{b,b'}$ coincide with the respective rate constants $k_{a,a'}^{b,b'}$ and the Gibbs energies $G(a,b)$ are the internal Gibbs energies $g(a,b)$ of the molecular states. In this case, equation (1.3.16) takes the form

$$\frac{k_{a,a'}^{b,b'}}{k_{a',a}^{b',b}} = \exp\left(\frac{g(a',b') - g(a,b)}{k_B T}\right). \tag{3.2.2}$$

Note that, for macromolecules the Gibbs energies $g(a,b)$ of internal states are different from the internal energies $\epsilon(a,b)$ of such states, because they additionally include entropic contributions and solvent effects.

The transitions that involve binding or release of a ligand should be treated separately. Suppose that a transition from (a,b) to (a',b') is accompanied by binding of a ligand and the ligand is released in the backward transition. Then the forward transition rate is proportional to the ligand concentration c, i.e. $w_{a',a}^{b',b} = k_{a',a}^{b',b} c$, whereas for the backward transition $w_{a,a'}^{b,b'} = k_{a,a'}^{b,b'}$. Moreover, the Gibbs energies in (3.2.1) include now contributions

from ligand particles, i.e. $G(a,b) = g(a,b) + \mu$, where μ is the chemical potential of the ligand. For the considered weak solutions, one has $\mu = \mu_0 + k_B T \ln c$. Substitution of these expressions into equation (3.2.1) yields

$$\frac{k^{b,b'}_{a,a'}}{k^{b',b}_{a',a}} = \exp\left(\frac{g(a',b') - g(a,b) - \mu_0}{k_B T}\right). \tag{3.2.3}$$

In this equation, the ligand can be either a substrate or a product if reverse binding of a product molecule takes place.

As shown by Schnakenberg [128], one can derive further identities by considering different pathways in a Markov network. Suppose that the chosen pathway represents a closed cycle Γ that involves only the internal states of the molecule without the events of ligand binding or release. Then, by using equation (3.2.2), one can show that the identity

$$\prod_{\Gamma} \frac{k^{b,b'}_{a,a'}}{k^{b',b}_{a',a}} = \exp\left(\sum_{\Gamma} \frac{g(a',b') - g(a,b)}{k_B T}\right) = 1 \tag{3.2.4}$$

holds, with the multiplication on the left side performed over all transitions that belong to the chosen cycle. This is the analogue of equation 1.3.15.

If the pathway Γ involves a conversion of substrate to a product or back, application of condition (3.2.3) leads to a modified identity. For tryptophan synthase, it has the form

$$\prod_{\Gamma} \frac{w^{b,b'}_{a,a'}}{w^{b',b}_{a',a}} = \exp\left(\frac{\mu(\mathrm{trp}) + \mu(\mathrm{G3P}) - \mu(\mathrm{ser}) - \mu(\mathrm{IGP})}{k_B T}\right) \tag{3.2.5}$$

if the pathway Γ leads from the bottom to the top empty states (empty,empty) in the Markov network in figure 2.6, i.e. if it corresponds to conversion of the two substrate molecules IGP and serine to the two product molecules G3P and tryptophan.

The detailed balance condition (3.2.1) and the Schnakenberg identities (3.2.4) and (3.2.5) can be used to check the thermodynamic consistency of a Markov network, to find missing rate constants of some transitions, and to determine Gibbs energies of different states. Particularly, in the Markov network of tryptophan synthase, there is a transition from the state $(4,4)$ to $(3,3)$ that corresponds to the channeling of indole from the β- to the α-site. This transition has never been observed experimentally and its rate constant could not be measured. This rate constant can however be determined, as explained below, by using the identity (3.2.5) and additional experimental data.

Kishore *et al.* were able to determine the difference of the Gibbs free energies between the product molecules (G3P and tryptophan) and substrate molecules (IGP and serine) by measuring the respective equilibrium concentrations [186]. Under standard conditions ($c_0(\mathrm{IGP}) = c_0(\mathrm{ser}) = c_0(\mathrm{G3P}) = c_0(\mathrm{trp}) = 1$ M), the difference of the chemical potentials $\mu_0(\mathrm{IGP}) + \mu_0(\mathrm{ser}) - \mu_0(\mathrm{G3P}) - \mu_0(\mathrm{trp})$ in tryptophan synthase is equal to 20.46 $k_B T$.

By using the identity (3.2.5) and the known value of Δq for tryptophan synthase, reverse channeling transition rate can be determined as $k^{3,4}_{3,4} = 4.55\ \mathrm{s}^{-1}$. This is indeed

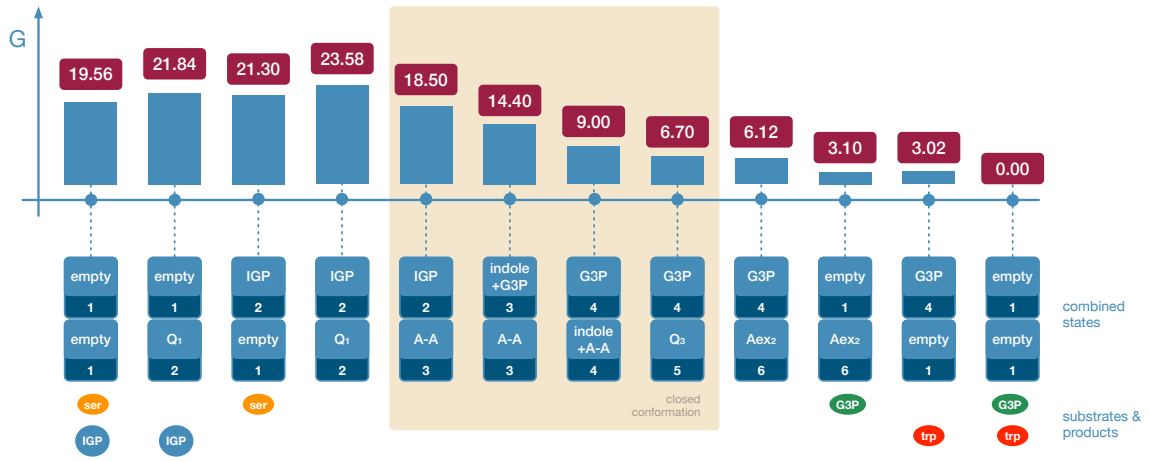


Figure 3.2: The Gibbs energy landscape along the main pathway of tryptophan synthase under physiological ligand concentrations. The Gibbs energies are given in units of k_BT. Physiological concentration values are chosen. In the states within the beige box, the molecular gates are closed and the enzyme is disconnected from the chemostats. Tryptophan is present inside the β-subunit in the state Aex_2.

much smaller than the measured rate $k_{4,3}^{4,3} = 1000\ \mathrm{s}^{-1}$ of the forward channeling transition. Therefore, the reverse channeling transitions should be very rare and this is why they have not been experimentally observed.

3.3 The Energy Landscape

With all transition rate constants on the fully reversible network 3.1, the detailed balance conditions (3.2.1) and (3.2.5) are now used to determine, by repeated application, the Gibbs energies $G(a, b)$ with respect to the Gibbs energy of a certain reference state.

The reference state corresponds to the free enzyme with two products (tryptophan and G3P) and its Gibbs energy is chosen as $G_{final} = 0$. In the initial state, the enzyme is free, there are two additional substrate molecules (serine and IGP) and the two product molecules (tryptophan and G3P) are missing. The Gibbs energy of the initial state is therefore $G_{initial} = \mu(\mathrm{IGP}) + \mu(\mathrm{ser}) - \mu(\mathrm{G3P}) - \mu(\mathrm{trp})$. It should be noted that it depends on the involved ligand concentrations c because $\mu = \mu_0 + k_BT \ln c$. It coincides with the amount of heat Δq released in one turnover cycle. The value above given $\Delta q = 20.46\ k_BT$ corresponds to the standard conditions $c_0(\mathrm{IGP}) = c_0(\mathrm{ser}) = c_0(\mathrm{G3P}) = c_0(\mathrm{trp}) = 1$ M. Recalculating this under the physiological concentrations (table 2.4) gives $G_{initial} = \Delta q = 19.56\ k_BT$.

There are also several states where one of the subunits is empty and the other subunit has a ligand bound to it. For example, the state (IGP, empty) has IGP bound to the α-subunit and no ligand in the β-subunit. The Gibbs energy of this state is $G(\mathrm{IGP,empty}) = g(\mathrm{IGP,empty}) - g_0 + \mu(\mathrm{ser}) - \mu(\mathrm{G3P}) - \mu(\mathrm{trp})$. It includes both the difference of the chemical potentials, depending on the concentrations, and the internal Gibbs energies $g(\mathrm{IGP,empty})$ and $g_0 = g(\mathrm{empty,empty})$ of the state (IGP, empty) and the

free state of the enzyme.

Finally, there are states where both subunits are occupied. For example, for the state (IGP, Q_1), we have $G(\text{IGP},Q_1) = g(\text{IGP},Q_1) - g_0 - \mu(\text{G3P}) - \mu(\text{trp})$. For the state (IGP, A-A), we have $G(\text{IGP,A-A}) = g(\text{IGP,A-A}) - g_0 - \mu(\text{G3P}) - \mu(\text{trp})$. Note that the difference $G(\text{IGP},Q_1) - G(\text{IGP,A-A}) = g(\text{IGP},Q_1) - g(\text{IGP,A-A})$ is determined only by the internal Gibbs energies of the states and is independent of ligand concentrations. This difference gives the amount of heat dissipated in the respective transition.

Figure 3.2 shows the Gibbs energy landscape of tryptophan synthase along its main pathway. After the binding of substrates requiring activation energies of 1.74 k_BT for IGP binding and 2.28 k_BT for serine binding, all transitions towards product formation are exergonic. The four catalytically important transitions (IGP,Q_1) $\rightleftharpoons$ (IGP,A-A) $\rightleftharpoons$ (indole+G3P,A-A) $\rightleftharpoons$ (G3P,indole+A-A) $\rightleftharpoons$ (G3P,Q_3) in the closed conformation of the enzyme are highly exergonic and accompanied by heat release in the range between 5.40 and 2.30 k_BT. The release of the products G3P and tryptophan is accompanied by the heat release of 3.10 and 3.02 k_BT, respectively.

3.4 Entropy Production and Flow

The theory of stochastic thermodynamics on fully reversible Markov networks arbitrarily far from equilibrium has been introduced in section 1.3. The central quantities are briefly restated here for the sake of readability and then calculated and interpreted for the fully reversible Markov network of tryptophan synthase under physiological conditions.

As before, the time evolution of the probability distribution $p(a, b; t)$ on the Markov network 3.1 is given by the master equation 3.1.3. The Shannon entropy at time t is defined as

$$S(t) = -\sum_{a,b} p(a, b; t) \ln p(a, b; t) \tag{3.4.1}$$

Its time derivative is

$$\frac{d}{dt}S = \frac{1}{2}\sum_{a,a',b,b'} J^{b,b'}_{a,a'} \ln \frac{p(a', b'; t)}{p(a, b; t)}. \tag{3.4.2}$$

It can be decomposed as

$$\frac{d}{dt}S = \sigma - h \tag{3.4.3}$$

into the difference of the entropy production σ inside the enzyme and of the net flow h of entropy *from* the enzyme, i.e. of the rate of entropy export by it, where

$$\sigma = \frac{1}{2}\sum_{x,x'} J_{x,x'} \ln \frac{w_{x,x'}p(x'; t)}{w_{x',x}p(x; t)} \tag{3.4.4}$$

and

$$h = \frac{1}{2} \sum_{x,x'} J_{x,x'} \ln \frac{w_{x,x'}}{w_{x',x}}. \tag{3.4.5}$$

dS/dt can be written as a sum of the contributions $s^{b,b'}_{a,a'}$ from each individual transition, i.e.

$$\frac{d}{dt}S = \frac{1}{2} \sum_{a,a',b,b'} s^{b,b'}_{a,a'}, \text{ with } s^{b,b'}_{a,a'} = J^{b,b'}_{a,a'} \ln \frac{p(a',b')}{p(a,b)}. \tag{3.4.6}$$

The same holds for the total entropy production σ and the rate of entropy export h

$$h = \frac{1}{2} \sum_{a,a',b,b'} h^{b,b'}_{a,a'};\ \sigma = \frac{1}{2} \sum_{a,a',b,b'} \sigma^{b,b'}_{a,a'} \tag{3.4.7}$$

where

$$h^{b,b'}_{a,a'} = J^{b,b'}_{a,a'} \ln \frac{w^{b,b'}_{a,a'}}{w^{b',b}_{a',a}}, \tag{3.4.8}$$

$$\sigma^{b,b'}_{a,a'} = J^{b,b'}_{a,a'} \ln \frac{w^{b,b'}_{a,a'} p(a',b')}{w^{b',b}_{a',a} p(a,b)}. \tag{3.4.9}$$

The properties and the physical meaning of $\sigma^{b,b'}_{a,a'}$, $h^{b,b'}_{a,a'}$ and $s^{b,b'}_{a,a'}$ are discussed in section 1.3.

Now, the Shannon entropy, the entropy production and entropy flow are calculated for the whole Markov network of tryptophan synthase (equations 3.4.1, 3.4.4 and 3.4.5) and for all transitions (equations 3.4.6, 3.4.8 and 3.4.9).

In the state of thermal equilibrium, all fluxes $J^{b,b'}_{a,a'}$ vanish and therefore according to equations 3.4.8 and 3.4.9 there are no transitions where entropy is produced or exported. Under physiological conditions, however, the enzyme tryptophan synthase operates far from thermal equilibrium, with the difference of Gibbs energies of 19.56 k_BT for one cycle. Thus, its operation is characterized by nonequilibrium steady-state. In the respective nonequilibrium steady-state with the stationary probability distribution $\bar{p}(a,b)$, the fluxes $\bar{J}^{b,b'}_{a,a'}$ do not vanish and therefore the transitions are accompanied by entropy production and entropy export. Because the entropy S is conserved in this state, $dS/dt = \sigma - h = 0$. Hence the total entropy production σ is counterbalanced by the entropy export h. Note that, although $dS/dt = 0$, the rates of entropy change $s^{b,b'}_{a,a'}$ for individual transitions are not zero even in the nonequilibrium steady-state.

The stationary probability distribution $\bar{p}(a,b)$ can be found by solving the master equation 3.1.3 in the nonequilibrium steady-state. Numerical values of the probabilities $\bar{p}(a,b)$ corresponding to all possible states are given in table B.6. Then, by using equation 1.3.2, the fluxes $\bar{J}^{b,b'}_{a,a'}$ can be determined. From equations 3.4.6, 3.4.8 and 3.4.9, the values of $\sigma^{b,b'}_{a,a'}$, $h^{b,b'}_{a,a'}$ and $s^{b,b'}_{a,a'}$ can be calculated afterwards.

The results are displayed in Figs. 3.3 and 3.4. The same network as in figure 3.1 is shown, but, for simplicity, only the numerical notations of the states are retained. Only the links between the states are shown because the transition directions are not important as the quantities $\sigma_{a,a'}^{b,b'}$ and $h_{a,a'}^{b,b'}$ are symmetrical, i.e. $\sigma_{a,a'}^{b,b'} = \sigma_{a',a}^{b',b}$ and $h_{a,a'}^{b,b'} = h_{a',a}^{b',b}$. For each link, the value of the quantities $\sigma_{a,a'}^{b,b'}$ or $h_{a,a'}^{b,b'}$ is indicated. Additionally, color coding is used.

Here and below, all numerical values for entropy and information are given in units of bits. One has 1 bit $= \ln 2 = 0.693$, because natural logarithms are used in the definition of the Shannon entropy.

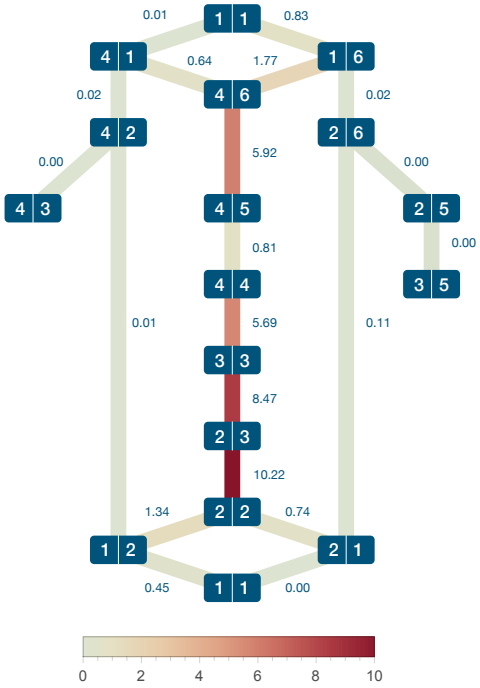


Figure 3.3: Entropy production in different transitions in the nonequilibrium steady-state. The values of entropy production are given in units of bit s^{-1} next to the links between the states. Additionally, color coding of the links according to the corresponding entropy production is used. The states are labeled according to table 2.3.

The rates of entropy or information change are given in bits per seconds. Alternatively, they can also be expressed by the respective amounts per a catalytic cycle. Note that the substrate conversion rate of the enzyme is equal to the probability flux $J^{channel}$ because each productive cycle includes this transition. The mean catalytic cycle time is the inverse of the substrate conversion rate. Under physiological concentrations the mean cycle time is 0.75 s. Tryptophan synthase is a slow molecular machine.

Figure 3.3 shows numerical values of entropy production for all individual transitions within the enzyme. The entropy is mostly produced along the main catalytic pathway. The highest entropy production (10.22 bit s^{-1}) is found for the allosterically activated transition $Q_1 \rightleftharpoons$ A-A in the β-site. In contrast to this, all transitions involving futile states (side branches of the network) have values of entropy production below 0.01 bit s^{-1}

per second. Ligand binding and release is characterized by entropy production below 1.78 bit s^{-1} per second.

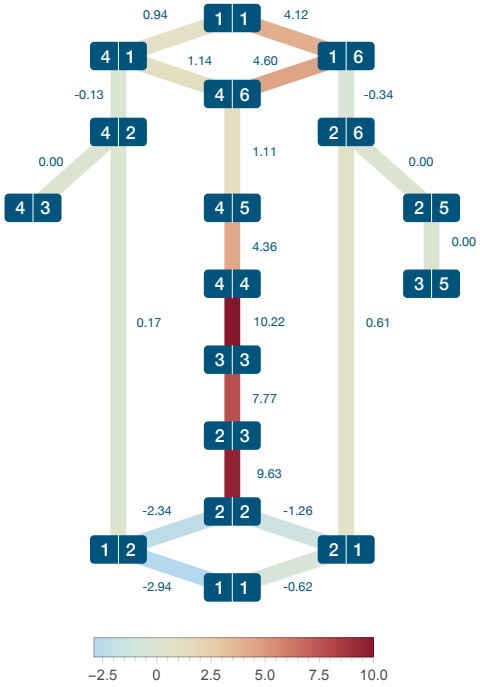


Figure 3.4: Rates of entropy export in individual transitions in tryptophan synthase. The same notations as in figure 3.3.

The values for entropy export are given in figure 3.4. The entropy export takes is maximal (between 7.77 bit s^{-1} and 10.22 bit s^{-1}) for the transitions (IGP,Q_1) $\rightleftharpoons$ (IGP,A-A) $\rightleftharpoons$ (indole+G3P,A-A) $\rightleftharpoons$ (G3P,indole+A-A) where most of the heat exchange with the environment takes place. All other transitions have absolute values smaller than 4.61 bit s^{-1}. Note that transition (G3P,Q_3) $\rightleftharpoons$ (G3P,Aex_2) has a small entropy export, but a high entropy production.

Because the rate of entropy change in a transition is given by the difference of entropy production and export, this rate can be found by subtracting the respective values in Figs. 3.3 and 3.4. Thus the transition (G3P,Q_3) $\rightleftharpoons$ (G3P,Aex_2) in the main catalytic pathway has the largest rate of entropy increase $s^{6,5}_{4,4} = 4.82$ bit s^{-1}. In contrast to this, channeling and the subsequent transition (G3P,indole+A-A) $\rightleftharpoons$ (G3P,Q_3) are accompanied by the net export of entropy at the rates $s^{4,4}_{3,3} = -4.53$ bit s^{-1} and $s^{5,4}_{4,4} = -3.55$ bit s^{-1}.

Using the computed rates of entropy production and export for individual transitions, total amounts for the whole enzyme per a turnover cycle can be obtained. Within a single catalytic cycle of tryptophan synthase, 27.79 bits of entropy are produced. The same amount of entropy is on the average exported by the enzyme per one cycle.

3.5 Discussion

In this study, methods of stochastic thermodynamics have been applied to characterize the operation of the channeling enzyme tryptophan synthase.

Using thermodynamic identities related to the detailed balance, the Gibbs energy landscape of this enzyme along its main catalytic pathway could be reconstructed from the experimental data. Under *in vivo* conditions, the cycle of this enzyme is driven by the Gibbs energy gradient of approximately 19.56 k_BT between its substrates and products. Thus, under physiological substrate and product concentrations, the enzyme operation is far from thermal equilibrium.

Inside the cycle of tryptophan synthase, only the first substrate binding transitions are thermally activated, with activation energies about 1 k_BT. All other transitions, including the events of product release, correspond to a decrease in the Gibbs energy. In particular, channeling is driven by the energy difference of 5.4 k_BT and does therefore not represent a diffusion process.

Because the enzyme operates far from equilibrium, entropy is persistently produced. It was found that 27.79 bits of entropy are produced and the same amount of entropy is exported, on the average, to the environment within one catalytic cycle. The distribution of entropy production over the Markov network is largely nonuniform.

Chapter 4

Information Exchange in Bipartite Systems

In this chapter, the analysis of information exchange and entropy production for bipartite Markov networks [22, 23, 24] is extended to systems that also have cross-transitions between the two subsystems. These results have been published in [25]. Resuming the argument of section 1.3.2, information theory enters the thermodynamics of stochastic processes when measurements of a subsystem if performed by another subsystem. This changes the free energy of the system being measured, but does neither influence its dynamics nor its current state - the measurement is non-interactive. This might seem contradictory to the gain in free energy at first glance. Yet, the free energy is a quantification of the work that can be potentially extracted from the system. The improved knowledge of the system's state increases the extractable work, because the respective experimental protocols can be adjusted according to the surplus of information. In contrast, the state of the measurement device is altered as a result of the measurement - after all the measurement device is a physical system and the storage of information necessitates a change of the physical state [188]. Note that this situation is asymmetrical: The evolution of the subsystem under measurement influences the dynamics of the measurement device, but not vice versa.

It is possible to symmetrize this situation by allowing the measurement to proceed in both directions, i.e. to allow both subsystems to measure each other, and by choosing a measurement device with its own internal dynamics (an ideal measurement device should of course be stable and not posses its own dynamics for a faithful storage of information, but this condition is relaxed with respect to the symmetrization). This leads naturally to the notion of a bipartite Markov network. Let $A \times B$ be a system composed of two subsystems A and B that have their own dynamics and perform measurements on each other. Denote the states of A and B by discrete variables $\{a|a \in A\}$ and $\{b|b \in B\}$ and the states of $A \times B$ by pairs $\{(a,b)|a \in A, b \in B\}$ and the transition probability rate for the transition from (a',b') to (a,b) as $w_{a,a'}^{b,b'}$. As discussed before, a measurement only changes the state of the measurement device, i.e. if A is measured by B, the corresponding transitions should be transitions from any state (a,b) to (a,b'), but not to (a',b') with $a \neq a'$. Analogously, transitions corresponding to the measurement of B by A should take place only between (a,b) and (a',b). Note that for a measurement of B by A, the transitions from (a,b) to (a',b) and from (a,b') to (a',b') can have different rates

$w^{b,b}_{a',a}$ and $w^{b',b'}_{a',a}$ although they correspond to the same transition of the A-subsystem. This dependence of the A-transition rates on the state of B is precisely the way in which A measures B: If $w^{b,b}_{a',a}$ and $w^{b',b'}_{a',a}$ were equal for all states b, b', then the transition between a to a' would not correspond to a measurement. If for each A-transition between a to a', $w^{b,b}_{a',a}$ and $w^{b',b'}_{a',a}$ were equal for all states b, b', then A could not measure B at all.

The internal dynamics of one subsystem is not directly linked to the internal dynamics of the other system and thus simultaneous transitions between (a, b) and $(a', b'), a \neq a', b \neq b'$ do not occur on a fine enough time scale. Thus, the same transitions that correspond to measurements at the same time correspond to the internal dynamics of each subsystem.

A bipartite Markov network is defined as a Markov-network corresponding to the situation just discussed, i.e. a Markov network on a product state space $A \times B$ such that the transition probability rates $w^{b,b'}_{a,a'}$ are zero whenever $a \neq a'$ and $b \neq b'$. In [23], a the theory is developed for such systems with the main result that the mutual information quantifies the information transfer between the two systems and that, in the steady state, the apparent entropy production (i.e. the entropy production in one subsystem determined if only the dynamics of this system is known and the dynamics of the other subsystem is inaccessible) within each subsystem is altered with respect to the real entropy production precisely by this information transfer. This allows, for example, to have an apparently negative entropy production in one subsystem in seeming contradiction to the second law. However, this is comes at the cost of a higher apparent entropy production in the other subsystem. The approach in the following section was motivated by ref. [23] and uses the same formalism and the same ideas. Therefore, a recap of [23] is unnecessary as it is a special case of the following.

4.1 General Formalism

Consider a system $A \times B$ composed of two subsystems A and B. The states of the system are labeled as (a, b). Assume that the bipartite transitions, i.e. the transitions of the form $(a, b) \rightleftharpoons (a, b')$ and $(a, b) \rightleftharpoons (a', b)$ that occur within one subsystem, have rates that can be affected by the state of the other subsystem. Moreover, in contrast to [23], also cross-transitions where the states of both subsystems become simultaneously changed, i.e. $(a, b) \rightleftharpoons (a', b')$ with $a \neq a'$ and $b \neq b'$, are also allowed. (For tryptophan synthase, there is one such transition and it corresponds to indole channeling.)

The evolution of the joint probability distribution $p(a, b; t)$ obeys the master equation

$$\frac{d}{dt}p(a, b; t) = \sum_{a',b'}[w^{b,b'}_{a,a'}p(a', b'; t) - w^{b',b}_{a',a}p(a, b; t)], \tag{4.1.1}$$

where $w^{b,b'}_{a,a'}$ denotes the transition rate from a state (a', b') to the state (a, b). Distinguishing between the regulatory and cross-transitions, one can write

$$w^{b,b'}_{a,a'} = \begin{cases} w^{b,b'}_{a} & \text{if } a = a' \\ w^{b}_{a,a'} & \text{if } b = b' \\ w^{b,b'}_{a,a'} & \text{if } a \neq a' \text{ and } b \neq b' \end{cases} \tag{4.1.2}$$

Moreover, probability fluxes are introduced as

$$J^{b}_{a,a'} = w^{b,b'}_{a,a'} p(a', b'; t) - w^{b',b}_{a',a} p(a, b; t) \text{ if } b = b', \tag{4.1.3}$$

$$J^{b,b'}_{a} = w^{b,b'}_{a,a'} p(a', b'; t) - w^{b',b}_{a',a} p(a, b; t) \text{ if } a = a', \tag{4.1.4}$$

$$J^{b,b'}_{a,a'} = w^{b,b'}_{a,a'} p(a', b'; t) - w^{b',b}_{a',a} p(a, b; t) \text{ if } a \neq a' \text{ and } b \neq b'. \tag{4.1.5}$$

The mutual information $i(a, b)$ for a pair of states (a, b) is defined as

$$i(a, b) = \ln \frac{p(a, b)}{p_A(a) p_B(b)}, \tag{4.1.6}$$

where $p_A(a) = \sum_b p(a, b)$ is the probability to find the subsystem A in the state a and $p_B(b) = \sum_a p(a, b)$. The average of $i(a, b)$ over all states (a, b) yields the mutual information I of the entire system

$$I = \sum_{a,b} p(a, b) \ln \frac{p(a, b)}{p_A(a) p_B(b)} = \sum_{a,b} p(a, b) i(a, b). \tag{4.1.7}$$

Its time derivative dI/dt can be written in the form

$$\frac{d}{dt} I = \frac{1}{2} \sum_{a,a'} f^{A}_{a,a'} + \frac{1}{2} \sum_{b,b'} f^{B}_{b,b'} + \frac{1}{2} \sum_{a,a',b,b'} f^{b,b'}_{a,a'}, \tag{4.1.8}$$

where the first two sums are taken over bipartite transitions in subsystems A or B and the last sum includes all cross-transitions in the considered system. One has

$$\begin{aligned} f^{A}_{a,a'} &= \sum_{b} J^{b}_{a,a'} \left[i(a, b) - i(a', b) \right] \\ &= \sum_{b} J^{b}_{a,a'} \ln \frac{p_B(b|a)}{p_B(b|a')}, \end{aligned} \tag{4.1.9}$$

$$\begin{aligned} f^{B}_{b,b'} &= \sum_{a} J^{b,b'}_{a} \left[i(a, b) - i(a, b') \right] \\ &= \sum_{a} J^{b,b'}_{a} \ln \frac{p_A(a|b)}{p_A(a|b')}, \end{aligned} \tag{4.1.10}$$

$$f^{b,b'}_{a,a'} = J^{b,b'}_{a,a'} \left[i(a, b) - i(a', b') \right]. \tag{4.1.11}$$

Here $p_A(a|b) = p(a, b)/p_B(b)$ is the conditional probability to find the A-system in state a if the B-system is in the state b and $p_B(b|a)$ is defined similarly.

Thus $f^{A}_{a,a'}$ yields the contribution to the total rate of change of mutual information due to the regulatory transition between a and a' that takes place in the subsystem A and is regulated by the subsystem B. A similar interpretation holds for $f^{B}_{b,b'}$. The term $f^{b,b'}_{a,a'}$ represents the contribution to the total rate of change of mutual information due to the cross-transition between (a, b) and (a', b'), with $a \neq a'$ and $b \neq b'$, that directly connects the two subsystems A and B.

Now the influence of the coupling through bipartite and cross-transitions on each of the entire subsystems A and B is derived. Therefore, consider the amount of entropy Σ^A produced per unit time in the transitions that change the state of the A subsystem. It is given by equation 3.4.9,

$$\Sigma^A = \frac{1}{2}\sum_{a,a',b} J^b_{a,a'} \ln \frac{w^b_{a,a'} p(a',b)}{w^b_{a',a} p(a,b)} + \\ + \frac{1}{2}\sum_{a\neq a', b\neq b'} J^{b,b'}_{a,a'} \ln \frac{w^{b,b'}_{a,a'} p(a',b')}{w^{b,b'}_{a,a'} p(a,b)}. \tag{4.1.12}$$

In a similar way, the amount of entropy Σ^B produced in the B subsystem can be found

$$\Sigma^B = \frac{1}{2}\sum_{a,b,b'} J^{b,b'}_{a} \ln \frac{w^{b,b'}_{a} p(a,b')}{w^{b',b}_{a} p(a,b)} + \\ + \frac{1}{2}\sum_{a\neq a', b\neq b'} J^{b,b'}_{a,a'} \ln \frac{w^{b,b'}_{a,a'} p(a',b')}{w^{b,b'}_{a,a'} p(a,b)}. \tag{4.1.13}$$

Suppose that the subsystem A is observed without the knowledge of the states of the subsystem B, i.e. there is no access to the joint probability distribution $p(a,b)$ and instead the probability distribution $p_A(a)$ in equation 4.1.12 is used. Proceeding in this way, the *apparent* entropy production σ^A assigned to the subsystem A is obtained

$$\sigma^A = \frac{1}{2}\sum_{a,a',b} J^b_{a,a'} \ln \frac{w^b_{a,a'} p_A(a')}{w^b_{a',a} p_A(a)} + \\ + \frac{1}{2}\sum_{a\neq a', b\neq b'} J^{b,b'}_{a,a'} \ln \frac{w^{b,b'}_{a,a'} p_A(a')}{w^{b',b}_{a',a} p_A(a)}. \tag{4.1.14}$$

Similarly, one obtains

$$\sigma^B = \frac{1}{2}\sum_{a,b,b'} J^{b,b'}_{a} \ln \frac{w^{b,b'}_{a} p_B(b')}{w^{b',b}_{a} p_B(b)} + \\ + \frac{1}{2}\sum_{a\neq a', b\neq b'} J^{b,b'}_{a,a'} \ln \frac{w^{b,b'}_{a,a'} p_B(b')}{w^{b',b}_{a',a} p_B(b)}. \tag{4.1.15}$$

The real entropy production rates Σ^A and Σ^B are always non-negative, whereas the apparent entropy production rates σ^A and σ^B can also be negative [22, 23, 24]. The influence on the entropy production of system A (respectively, B) through coupling to the whole system is then given by the difference between the apparent and total entropy production. Thus one defines

$$F^A = \sigma^A - \Sigma^A \tag{4.1.16}$$
$$F^B = \sigma^B - \Sigma^B. \tag{4.1.17}$$

Substituting equations 4.1.12 to 4.1.24 gives

$$\begin{aligned} F^A = &\frac{1}{2}\sum_{a,a',b} J^b_{a,a'} \ln \frac{p_B(b|a)}{p_B(b|a')} + \\ &\frac{1}{2}\sum_{a\neq a', b\neq b'} J^{b,b'}_{a,a'} \ln \frac{p_B(b|a)}{p_B(b'|a')}, \end{aligned} \tag{4.1.18}$$

$$\begin{aligned} F^B = &\frac{1}{2}\sum_{b,b',a} J^{b,b'}_{a} \ln \frac{p_A(a|b)}{p_A(a|b')} + \\ &+\frac{1}{2}\sum_{a\neq a', b\neq b'} J^{b,b'}_{a,a'} \ln \frac{p_A(a|b)}{p_A(a'|b')}. \end{aligned} \tag{4.1.19}$$

Note that F^A and F^B have contributions from terms $f^A_{a,a'}$ and $f^B_{b,b'}$ defined in equations 4.1.9 and 4.1.10 and used in the splitting of dI/dt in equation 4.1.8. In addition, they also include cross-terms originating from non-bipartite transitions.
Using F^A and F^B, equation 4.1.8 for the rate of change of mutual information can be written as

$$\frac{d}{dt}I = F^A + F^B + F^{cross} \tag{4.1.20}$$

where the quantity

$$F^{cross} = \frac{1}{2}\sum_{a\neq a', b\neq b'} J^{b,b'}_{a,a'} \ln \frac{p(a',b')}{p(a,b)}. \tag{4.1.21}$$

is introduced. Note that the expression 4.1.21 for F^{cross} can be also formulated as

$$F^{cross} = \sum_{a\neq a', b\neq b'} s^{b,b'}_{a,a'} \tag{4.1.22}$$

where $s^{b,b'}_{a,a'}$ is the Shannon entropy produced in the cross-transition from (a,b) to (a',b'). Using the non-negativity of Σ^A and Σ^B, one arrives at the second law-like inequalities

$$\Sigma^A = \sigma^A - F^A \geq 0, \tag{4.1.23}$$
$$\Sigma^B = \sigma^B - F^B \geq 0, \tag{4.1.24}$$

where F^A and F^B are related by the change of mutual information and the rate of Shannon entropy in the cross-transitions according to equation 4.1.20.

The equations 4.1.23 and 4.1.24 are the same as previously derived for completely bipartite systems where two subsystems were coupled by regulatory transitions, but no

cross-transitions were allowed [22, 23, 24]. In the absence of cross-transitions, the original framework[22, 23, 24] is recovered. Now, these inequalities have been generally derived for the systems where both regulatory and cross-transitions directly connecting the subsystems can take place. Such generalization is only possible if the definitions 4.1.18 and 4.1.19 are employed. Once the inequalities have been established, the same interpretation as in refs. [22, 23, 24] can be used.

4.2 Information Exchange in Tryptophan Synthase

There is a complex pattern of allosteric interactions between the two subunits of tryptophan synthase. Additionally, one transition that corresponds to indole channeling and affects simultaneously both subunits takes place. The allosteric cross-regulations and channeling lead to the development of correlations between the internal states of the subunits. In chapter 2, the presence of correlations has been demonstrated by computing the Pearson correlation coefficients for all possible pairs of states. In this section, the concept of mutual information will be employed to further quantify the effects of allosteric cross-regulation and channeling based on the theoretical framework presented in the previous section. The kinetic model from chapter 3 is used here.

The mutual information $i(a,b)$ between the states a and b of the two subunits is defined by equation 4.1.6, where $p_\alpha(a) = \sum_{b=1}^{6} p(a,b)$ and $p_\beta(b) = \sum_{a=1}^{4} p(a,b)$ are the probability distributions for the states of α- and β-subunits. $i(a,b)$ quantifies correlations between the states a of the α-subunit and b of the β-subunit, it vanishes if these states are statistically independent, i.e. if $p(a,b) = p_\alpha(a)p_\beta(b)$. If it is negative, anti-correlations between the states are present.

The values $i(a,b)$ under physiological conditions are shown in figure 4.1 for all states (a,b). High correlations (2.39 and 2.20 bits) are found between the states (G3P,indole+A-A) and (G3P,Q_3) after indole channeling and after the indole reaction at the β-site in the main pathway. This agrees with the previous analysis using the Pearson correlation coefficients [21]. As a result of channeling, both subunits simultaneously arrive at the state (G3P,indole+A-A) and high positive correlations are characteristic for it. On the other hand, anticorrelation (-1.04 bits) in the state (IGP,A-A) before channeling is present. This is an effect of allosteric interactions: when the β-subunit is in the state A-A, the cleavage of IGP into G3P and indole is blocked when the β-subunit is in the state Q_1, but it is possible in the state A-A.

The statistical average of $i(a,b)$ over all pair states (a,b) yields the mutual information I of the whole system

$$I = \sum_{a=1}^{4}\sum_{b=1}^{6} p(a,b)\ln\frac{p(a,b)}{p_\alpha(a)p_\beta(b)} = \sum_{a=1}^{4}\sum_{b=1}^{6} p(a,b)i(a,b). \tag{4.2.1}$$

This property is positive and it characterizes the strength of statistical correlations between the α- and β-subunits. For tryptophan synthase under physiological conditions one obtains $I = 0.49$ bit.

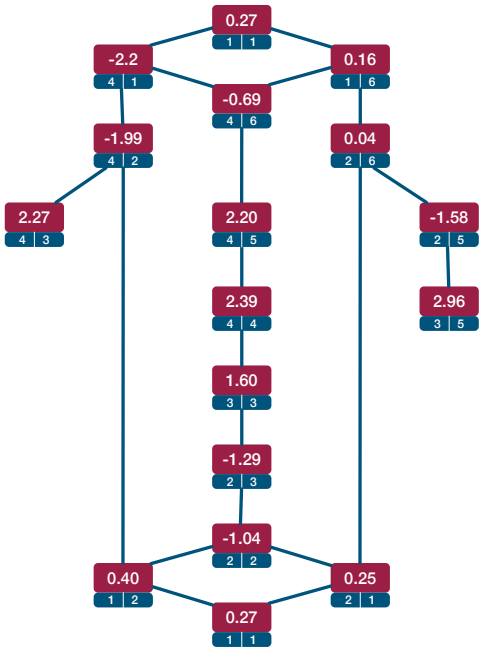


Figure 4.1: Correlations $i(a, b)$ in units of bits for different states a and b.

By equation 4.1.8, the rate of mutual information change for the entire system is

$$\frac{d}{dt}I = \frac{1}{2}\sum_{a,a'=1}^{4}{}' f^{\alpha}_{a,a'} + \frac{1}{2}\sum_{b,b'=1}^{6}{}' f^{\beta}_{b,b'} + f^{channel}. \tag{4.2.2}$$

Here, the sums exclude the forward and backward channeling transitions and

$$f^{\alpha}_{a,a'} = \sum_{b} J^{b}_{a,a'} \left[i(a,b) - i(a',b) \right], \tag{4.2.3}$$

$$f^{\beta}_{b,b'} = \sum_{a} J^{b,b'}_{a} \left[i(a,b) - i(a,b') \right], \tag{4.2.4}$$

$$f^{channel} = J^{channel} \left[i(4,4) - i(3,3) \right]. \tag{4.2.5}$$

Note that in a steady state $dI/dt = 0$ and therefore the terms 4.2.3 - 4.2.5 satisfy one additional constraint. Moreover, the terms $f^{\alpha}_{a,a'}$ and $f^{\beta}_{b,b'}$ do not depend on the choice of a direction for the transitions between a and a' or between b and b'. The quantity $f^{\alpha}_{a,a'}$ gives the contribution by the transition between the states a and a' in the α-subunit to the rate of change of the total mutual information of the system; this contribution is averaged over all possible regulatory states of the subunit β. A similar interpretation holds for the quantity $f^{\beta}_{b,b'}$.

By solving the master equation under physiological concentrations, the steady state probabilities $\bar{p}(\alpha, \beta)$ are obtained. Substituting them into equation 1.3.2 and into the equations 4.2.3 - 4.2.5 yields the values for $f^{\alpha}_{a,a'}$, $f^{\beta}_{b,b'}$ and $f^{channel}$.

Figure 4.2 shows how the generation (or loss) of mutual information is distributed over the network. Mutual information is generated in three transitions in the α-subunit. Its

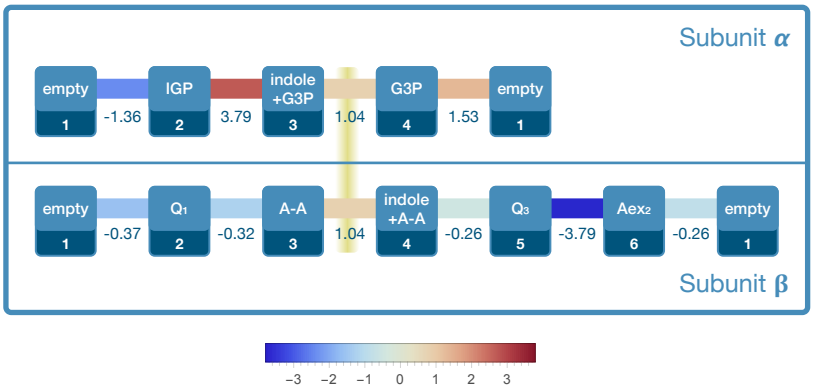


Figure 4.2: Rates of change of mutual information in units of bits per second for the transitions within α- and β-subunits and for the channeling transition.

highest generation rate is 3.79 bit s^{-1} in the transition (IGP $\rightleftharpoons$ indole+G3P) preceding channeling. The channeling transition itself generates mutual information at a smaller rate (1.04 bit s^{-1}). All transitions in the β-subunit are accompanied by mutual information loss with the highest rate (-3.79 bit s^{-1}) achieved in the transition immediately after channeling ($Q_3 \rightleftharpoons Aex_2$).

Furthermore, information interactions between entire subunits can also be discussed. To do this, the rate of change of mutual information is written as

$$\frac{d}{dt}I = F^{\alpha} + F^{\beta} + F^{channel} \tag{4.2.6}$$

where

$$F^{\alpha} = \frac{1}{2}\sum_{a,a',b} f^{\alpha}_{a,a'} + f^{\alpha}_{channel}, \tag{4.2.7}$$

$$F^{\beta} = \frac{1}{2}\sum_{a,b,b'} f^{\beta}_{b,b'} + f^{\beta}_{channel}. \tag{4.2.8}$$

Here, the rate of generation of mutual information in the channel $f^{channel}$ was divided, given by equation 4.2.5 into three parts, i.e. $f^{channel} = F^{channel} + f^{\alpha}_{channel} + f^{\beta}_{channel}$, where

$$F^{channel} = J^{channel} \ln \frac{p(3,3)}{p(4,4)}, \tag{4.2.9}$$

$$f^{\alpha}_{channel} = J^{channel} \ln \frac{p_{\alpha}(3)p(4,4)}{p_{\alpha}(4)p(3,3)}, \tag{4.2.10}$$

$$f^{\beta}_{channel} = J^{channel} \ln \frac{p_{\beta}(3)p(4,4)}{p_{\beta}(4)p(3,3)}. \tag{4.2.11}$$

Thus the rates of mutual information change in α- and β-subunits include now contributions $f^{\alpha}_{channel}$ and $f^{\beta}_{channel}$ from the channeling transition. The advantage of this definition is that, as shown in the previous section, the important thermodynamic inequalities 4.1.23 and 4.1.24 for the entropy production in both subunits become then

satisfied.

In a steady state, dI/dt vanishes and thus $F^{\alpha} + F^{\beta} + F^{channel} = 0$. If the channeling was absent, one would have had $F^{\alpha} = -F^{\beta}$. In this case, the mutual information generated in one subunit would have been completely consumed in the other subunit, cf. [22, 23, 24]. Because $F^{channel} \neq 0$, this is, however, no longer valid. Some mutual information for the entire enzyme is additionally generated in the channeling transition involving simultaneously both subunits.

The values for F^{α}, F^{β} and $F^{channel}$ under physiological concentrations have been computed. They all have the same order of magnitude. The mutual information $F^{channel} = -4.53$ bit s^{-1} generated per unit time by the transition corresponding to indole channeling flows to both subunits where is consumed at the rates of $F^{\alpha} = 3.09$ bit s^{-1} and $F^{\beta} = 1.42$ bit s^{-1}. Note that F^{β} is positive whereas all contributions $f^{\beta}_{b,b'}$ from individual transitions in the β-subunit are negative. This is an effect of the large contributions from the cross term $f^{\beta}_{channel} = 6.43$ bit s^{-1} (whereas $f^{\alpha}_{channel} = -0.86$ bit s^{-1}).

4.3 Discussion

Information interactions between the two catalytic subunits of the enzyme have been analyzed. Both the allosteric interactions between the subunits and the channeling of an intermediate product from one of them to another contribute to the change of mutual information. Thus, the previously existing theory [22, 23, 24] had to be generalized to the situations where, in addition to regulatory interactions between the subsystems, the transitions simultaneously changing the states of both of them can also take place. Mutual information is generated both in α- and β-subunits at the rates 3.09 and 1.49 bits per second. This mutual information is consumed in the channeling transition so that the balance is maintained. Moreover, contributions from individual allosterically regulated transitions in each of the subunits to the total mutual information change were determined.

Thus, it was demonstrated that, through the use of stochastic thermodynamics, a rich quantitative characterization of the nonequilibrium operation of an enzyme can be produced. It would be interesting to perform analogous investigations for other enzymes with several catalytic subunits. Such further investigations can clarify the connections between various thermodynamic properties of such nanomachines and the aspects of the chemical function of the enzymes.

Chapter 5

Semigroup Models for Reaction Networks

In this chapter, a new class of semigroup models for catalytic reaction systems (CRS) is presented. CRS are representations of chemical reaction networks with emphasis on the catalytic function of certain chemicals that are themselves part of the network. They have been introduced by Hordijk and Steel [189] as a generalization of Kauffman's autocatalytic sets [190] while studying the occurrence of self-sustaining subnetworks (called RAF sets in the parlance of CRS). Classically, chemical reaction networks are described by differential equations for the time evolution of concentrations of chemical species. Within this framework, it is not clear how to formally distinguish between metabolites and catalytically active enzymes. However, the concept of enzyme function has consolidated in biological sciences and was suggested by philosophers to be included in the quantitative natural sciences [191]. In section 5.1.1, the CRS formalism is motivated and introduced formally.

Within the semigroup formalism, the notion of enzyme function is extended to the successive and joint functions of arbitrary subsystems on the whole reaction system. In section 5.2, semigroup models for arbitrary CRS are constructed, their basic properties are discussed and the function of a subsystem is introduced. Section 5.3 extends the construction to CRS with food set. It is then shown that the maximal function of the CRS produces the CRS from the food set if and only if the system has the RAF property. A corollary is that the maximal function acting on the food set contains the maximal RAF set. In particular, if the semigroup is nilpotent, the CRS has no RAF sets. This is an important statement, because the vast majority of semigroups are nilpotent and the semigroups corresponding to RAF sets are thus located in the narrow class of non-nilpotent semigroups.

In section 5.4, a discrete dynamics is defined on the power set of the set of all chemicals. It is shown that dynamics has a fixed point if its initial condition is the whole set of chemicals. Moreover, this fixed point contains the maximal RAF set. Combining the methods from sections 5.3 and 5.4, it is possible to identify the maximal RAF set of any CRS (theorem 5.5.5). This is a main result of this chapter, because the identification of the maximal RAF of a CRS is a major challenge and receives a lot of attention in the literature. The section ends with a remark about the connection between the CRS formalism and the general formalism of chemical reaction networks (CRN). It is sketched

how a CRS can be rewritten as a CRN and thus the theory of thermodynamics of CRN is made accessible for CRS. In particular, this allows to exclude CRS that are not thermodynamically consistent.

The main motivation to use an algebraic formalism is the possibility of an algebraic coarse-graining procedure via quotient structures. Taking the quotient of a semigroup can be thought of as lumping together elements of the semigroup in such a way that the original semigroup operation naturally descends to an operation between the lumped states. The possible quotients are determined by the lattice of congruences that captures all algebraically allowed coarse-graining procedures. This motivation is further discussed in section 5.1.2 together with the formal definition of congruences and quotients of semigroups. In section 5.6, the application to the constructed semigroup models is demonstrated by the construction of two biologically relevant families of congruences. The first construction is a congruence on the subsemigroup of constant functions and reveals the organization of metabolic pathways within the CRS. The second construction is a family of congruences that leads to a rather unusual coarse-graining procedure. The network is covered with local patches in such a way that the local information about the network is fully retained, but the environment of each patch is no longer resolved. Whereas classical coarse-graining procedures would fix a particular local patch and delete detailed information about its environment, the algebraic approach keeps the structure of all local patches and even allows the interaction of functions within distinct patches.

The text uses a mathematically flavored language to avoid semantical ambiguities. Some definitions and theorems from semigroup theory are given in the introductory section 5.1.2 and some are included in the main text for the sake of better readability. They are then marked with an asterisk (*). A self-contained presentation of the concepts can be found in [192].

5.1 Motivation

5.1.1 Self-Sustaining Reaction Networks

Self-sustaining reaction networks form the basis of a class of theories for the origins of life based on the *cells first* hypothesis as advocated by Oparin [193], Dyson [194] and many others [195, 196, 197]. A self-sustaining reaction network is a reaction network that is able to generate all its substances from a given set of externally supplied chemicals (called food in the literature, c.f. [189, 198]). Its reactions are catalyzed and all catalysts are themselves part of the network. The main idea of the *cells first* hypothesis theories is based on the observation that micelles can form rather easily under prebiotic conditions. Such micelles enclose certain chemicals - in most cases just water and some other small molecules. Every now and then, they should contain molecules that are able to react with each other, i.e. that form a chemical reaction network. A reaction network with autocatalytic properties can possibly contain a self-sustaining subnetwork. A self-sustaining network has the inherent ability to replicate itself and grow. If its growth is supplemented with growth and division of micelles, this system is a potential candidate for primordial cells.

The main modern example of a self-sustaining network is the whole reaction network of an organism: The externally supplied food sources are carbohydrates and minerals. They are transformed into proteins, RNA, DNA and structural elements such as the cytoskeleton or cell membranes. There is a two-level system of "catalysts": Proteins are the direct chemical catalysts enabling the transformations on the network, whereas the formation of proteins is controlled (catalyzed in the language of reaction networks) by DNA via RNA. Including metabolites, proteins and DNA into one large reaction network makes it self-sustaining whereas subnetworks excluding any of these species are generally not self-sustaining.

From the author's point of view, the condition of self-sustainability is so crucial to the reaction networks of biological systems that any formalism attempting to model such networks should at least in principle be able to distinguish between reaction networks that can sustain themselves and those that cannot. One possibility towards suitable formalisms is to try to capture the organization of a reaction network. Both this viewpoint and the method of resolution have been advanced by prominent scientists.

In the 1940ies and 50ies, von Neumann was pioneering the development of automata theory alongside Turing, Church, Shannon and many others. The theory describes the organization and possible logical operations performed by a computing machine [199]. Using this new framework, von Neumann constructed a self-replicating automaton with the goal of modeling a living system [200]. However, he noted that the theory could only be complete if it was linked to thermodynamics thereby making his construction falsifiable under the three laws of thermodynamics [201]. The biophysicist Rashevsky, who is arguably one of founding fathers of mathematical biology, spent many years of his career working very successfully on models of partial processes in organisms such as intracellular oxygen diffusion [202], nerve excitation [203], or cell polarity [204]. Yet, later he concluded that such models merely capture subsystems of a living being without any relation to the whole organism. In particular, his models would still remain the same if the organism died. Therefore he suggested to use more abstract mathematical methods such as topology to capture the organization of an organism as a whole [205]. This approach was termed *relational biology* [206]. Its main focus was to capture the structure of interactions between the parts of an organism. The actual physical material forming the organism was seen as one possible realization of a relational structure. Rashevsky's student Rosen continued to work in this direction. He used the language of category theory to describe organizational structures that were self-referential [207, 208]. However, he was not able to link his formalism to actual physical phenomena.

A more chemical approach was given by Stuart Kauffman in 1986, when he introduced a binary polymer model to study the emergence of self-sustaining reaction networks [190]. In this model, two molecules a and b supplied from the environment are able to form linear polymers (represented by strings of a and b). The possible reactions are cleavage and fusion of polymers. This yields a reaction network. It is assumed that each polymer has a certain probability of catalyzing a reaction within the network. A reaction network of level N is a binary polymer network where the length of the polymers is at most N. The main result of Kauffman is the almost certain emergence of self-sustaining reaction networks for high enough level. A generalization of Kauffman's model under the name

of RAF networks was introduced by Hordijk and Steel [189]. This notion is also focused on the catalytic interactions between chemicals in a reaction network, but replaces the binary polymers with arbitrary chemicals. RAF networks are defined as a special case of catalytic reaction systems (CRS). The definitions given here follow [189].

Definition 5.1.1. A *catalytic reaction system* (CRS) is a triple (X, R, C), where X is a finite discrete set of chemicals, R is the finite set of reactions $r : X \to \mathbb{Z}$ and $C \subset X \times R$ is a set of reactions catalyzed by chemicals of X. For any pair $(x, r) \in C$, the reaction r is said to be catalyzed by x.

Definition 5.1.2. A *subnetwork* (X', R', C') of a CRS (X, R, C) is given by the subset $X' \subset X$ with the maximal possible sets of reactions and catalyzed reactions:

$$R' = \{r|_{X'} \text{ such that } r \in R, dom(r) \subset X', ran(r) \cap X' \neq \emptyset\},$$

where $r|_{X'}$ denotes the restriction of $r : X \to \mathbb{Z}$ to X' and

$$C' = \{(x, r|_{X'}) \text{ such that } \exists (x, r) \in C \text{ with } r|_{X'} \in R' \text{ and } x \in X'\}.$$

It is possible to have a some reaction r included in R, but not its reverse $-r$. This can be justified by the fact that many reactions proceed along a chemical potential gradient and are therefore essentially irreversible. One example is the reaction catalyzed the enzyme tryptophan synthase presented in the previous chapters.

Giving a reaction in the form $r : X \to \mathbb{Z}$ as above is equivalent to the usual notation

$$a_1A_1 + a_2A_2 + ... + a_nA_n \to b_1B_1 + b_2B_2 + ... + b_mB_m,$$

where $a_i, b_j \in \mathbb{N}$ and $A_i, B_j \in X$, $A_i \neq B_j$ for $i = 1, ..., n$ and $j = 1, ..., m$ via

$$r(x) = \begin{cases} -a_i, & \text{if } x = A_i \\ b_j, & \text{if } x = B_j \\ 0, & \text{else.} \end{cases}$$

The notation $r : X \to \mathbb{Z}$ allows the notion of linear combinations $(\sum_i \mu_i r_i) : X \to \mathbb{Z}, \mu_i \in \mathbb{Z}$ of reactions $\{r_i\}_{i \in I}$ via

$$(\sum_i \mu_i r_i)(x) := \sum_i \mu_i r_i(x)$$

and will therefore be used for notational convenience. It is useful to define the domain $dom(r)$ and range $ran(r)$ of a reaction as

$$dom(r) = \{x \in X, r(x) < 0\}$$

and

$$ran(r) = \{x \in X, r(x) > 0\}.$$

Following [209] a CRS can be graphically represented by a graph with two kinds of vertices and two kinds of directed edges. As an example, consider the graph in figure 5.1. The solid disks correspond to the chemicals in X and the circles corresponds to

reactions in R. The chemicals participating in a reaction are shown by solid arrows. If the reaction is catalyzed by some chemical, this is indicated by a dashed arrow. Usually, the stochiometry of a reaction is not explicitly shown in the graph.

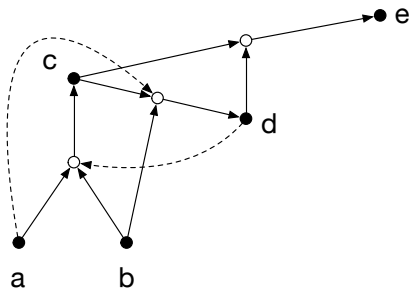


Figure 5.1: Example of a graphical representation of a CRS. The CRS consists of five chemicals $X = \{a, b, c, d, e\}$ and three reactions $a + b \to c$, $c + b \to d$ and $c + d \to e$. The first two reactions are catalyzed by d and a, respectively, whereas the last reaction is not catalyzed.

Definition 5.1.3. A *reflexive autocatalytic network* (RA network) is a CRS (X, R, C), such that each reaction $r \in R$ is catalyzed by some chemical $x \in X$, or, equivalently, if the natural projection $C \to R$ is surjective. The CRS (X, R, C) is said to posses the RA property.

Definition 5.1.4. A *CRS with food set* F is a quadruple (X, R, C, F), where (X, R, C) is a CRS and $F \subset X$. A *subnetwork of a CRS* (X, R, C, F) *with food set* F is a CRS with food set (X', R', C', F) such that $F \subset X'$ and (X', R', C') is a subnetwork of (X, R, C) by definition 5.1.2.

Definition 5.1.5. A *food-generated network* (F network) is a CRS with food set such that each $x \in X$ is generated by some sequence of reactions from F. The CRS (X, R, C) is said to be generated from the food set F. More precisely, (X, R, C) is generated from F if the following two conditions are satisfied:

(F1) For every $x \in X$ there is a finite index set I such that the linear combination $r := (\sum_{i \in I} \mu_i r_i), \mu_i \in \mathbb{N}$ of reactions $\{r_i\}_{i \in I} \subset R$ satisfies $x \in ran(r)$ and $dom(r) \subset F$ and the index set I satisfies the condition:

(F2) There is a partition of I

$$I = \coprod_{j=1}^{n} I_j$$

and reactions $\tilde{r}_j := (\sum_{i \in I_j} \mu_i r_i), j = 1, ..., n$ such that $dom(\tilde{r}_1) \subset F$ and $dom(\tilde{r}_{j+1}) \subset \cup_{k=1}^{j} ran(\tilde{r}_k)$ for $j = 1, ..., n-1$.

Remark 5.1.6. Intuitively, condition (F1) is enough to capture the notion of generation from a food set. However, condition (F2) makes the definition given here equivalent to

the original definition given by Hordijk and Steel [189]. It turns out to make a crucial difference between a RAF set and more general self-sustaining networks. This is discussed in remark 5.5.6 after the study of semigroup models and their connection to RAF sets and self-sustaining networks.

Remark 5.1.7. Each CRS (X, R, C) can be made into an F network by taking $F = X$. Due to the finiteness of X there exist minimal (not necessarily unique) food sets F for every CRS making it an F network.

Remark 5.1.8. In [189], a RAF network is defined as follows: *A RAF network is an F network* (X, R, C, F) *where* (X, R, C) *is RA.* This definition requires all possible reactions between chemicals in the food set F to be catalyzed. This is redundant, because these chemicals are supplied from the environment. Therefore the author prefers to use a slightly modified definition of a RAF network taking this minor detail into account (definition 5.3.2). Otherwise, the author's definition agrees with the definition given above.

The definition of a RAF network captures the intuitive notion of a self-sustaining chemical reaction network. The RA property allows each chemical to be formed by reactions catalyzed by the network itself and the generation from a food set implies that every chemical in the network can be regenerated from resources taken up from the environment.

Example 5.1.9. The subnetwork shown in figure 5.1 given by $\{a, b, c, d\}$ is RA, because all its reactions are catalyzed. Choosing the food set $F = \{a, b\}$ makes it into a RAF network as all chemicals in the network are generated from F. However, for $F = \{a\}$ the network is no longer RAF, because b cannot be generated from the food set. The RA property is a property of the network (X, R, C), whereas the F property is not inherent to the network, but depends on the choice of food set.

5.1.2 Coarse-Graining via Congruences

One benefit of algebraic models and the main motivation for this work is their natural hierarchy of substructures and quotient structures. As has been discussed in the thesis introduction, one important characteristic of biological systems are processes taking place on many length and time scales and an associated hierarchy of structures and interactions between them. However, current approaches for the transitions from a lower scale to higher scales crucially rely on a time scale separation. Even the transition from a given scale to one higher scale by integrating out fast degrees of freedom can be technically very demanding [210]. The use of algebraic structures is an attempt to circumvent these difficulties and try to perform a "coarse-graining in function" by taking quotients (section 5.6). An algebraic quotient groups together classes of elements in a way such that the classes are compatible under some given algebraic operations. This gives the quotient structure the algebraic type of the original structure. As an example without biological interpretation, consider the subgroups and quotients of $\mathbb{Z}$:

Example 5.1.10. The integers $\mathbb{Z} = \{0, \pm 1, \pm 2, ...\}$ form a commutative group under addition. Each subgroup of $\mathbb{Z}$ is of the form $n\mathbb{Z} = \{0, \pm n, \pm 2n, ...\}, n \in \mathbb{N}$. The subgroups $n\mathbb{Z}$ form a hierarchy fully determined by divisibility of the natural numbers, i.e. $m\mathbb{Z}$ is a subgroup of $n\mathbb{Z}$ if and only if n divides m. One writes $m\mathbb{Z} < n\mathbb{Z}$. If $m\mathbb{Z} < n\mathbb{Z}$, then the question about proper subgroups between $m\mathbb{Z}$ and $n\mathbb{Z}$ is determined by the quotient m/n:

There exist proper subgroups if and only if m/n is not a prime number. The resulting hierarchy of subgroups of $\mathbb{Z}$ is sketched in figure 5.2A.

Each subgroup $n\mathbb{Z}$ yields a quotient group $\mathbb{Z}/n\mathbb{Z} = \{\bar{0}, \bar{1}, ..., \overline{n-1}\}$ - the group of residue classes modulo n. As sets, the residue classes $\bar{i}$ are cosets $i + n\mathbb{Z} = \{..., i-2n, i-n, i, i+n, ...\}$. $\mathbb{Z}/n\mathbb{Z}$ inherits the addition from $\mathbb{Z}$, i.e.

$$\bar{i} + \bar{j} = \begin{cases} \overline{i+j} & \text{if } i+j < n \\ \overline{i+j-n} & \text{if } i+j \geq n. \end{cases}$$

Each quotient $\mathbb{Z}/m\mathbb{Z}$ sees the part of the lattice above the subgroup $m\mathbb{Z}$ and forgets the rest of it as shown in figure 5.2B for the quotient $\mathbb{Z}/60\mathbb{Z}$. The normal subgroups containing $m\mathbb{Z}$, i.e. the groups $n\mathbb{Z}$ such that $n|m$, become quotient groups $n\mathbb{Z}/m\mathbb{Z} \cong \mathbb{Z}/(m/n)\mathbb{Z}$. They can be used to take further quotients of the lattice as shown in figure 5.2C. Throughout this procedure, the addition defined on $\mathbb{Z}$ descends to a well-defined addition on all the quotients.

For the group $\mathbb{Z}$ (and any group) the quotients are in one-to-one correspondence with its normal subgroups and therefore the characterization of quotients given here is complete. However, for general algebras, the subalgebras do not determine all possible quotients. The appropriate notion is the notion of congruence relation. Congruences can be defined for any type of algebra [211], but the exposition here will focus on semigroups. This material is presented in greater detail in [192], chapter I.

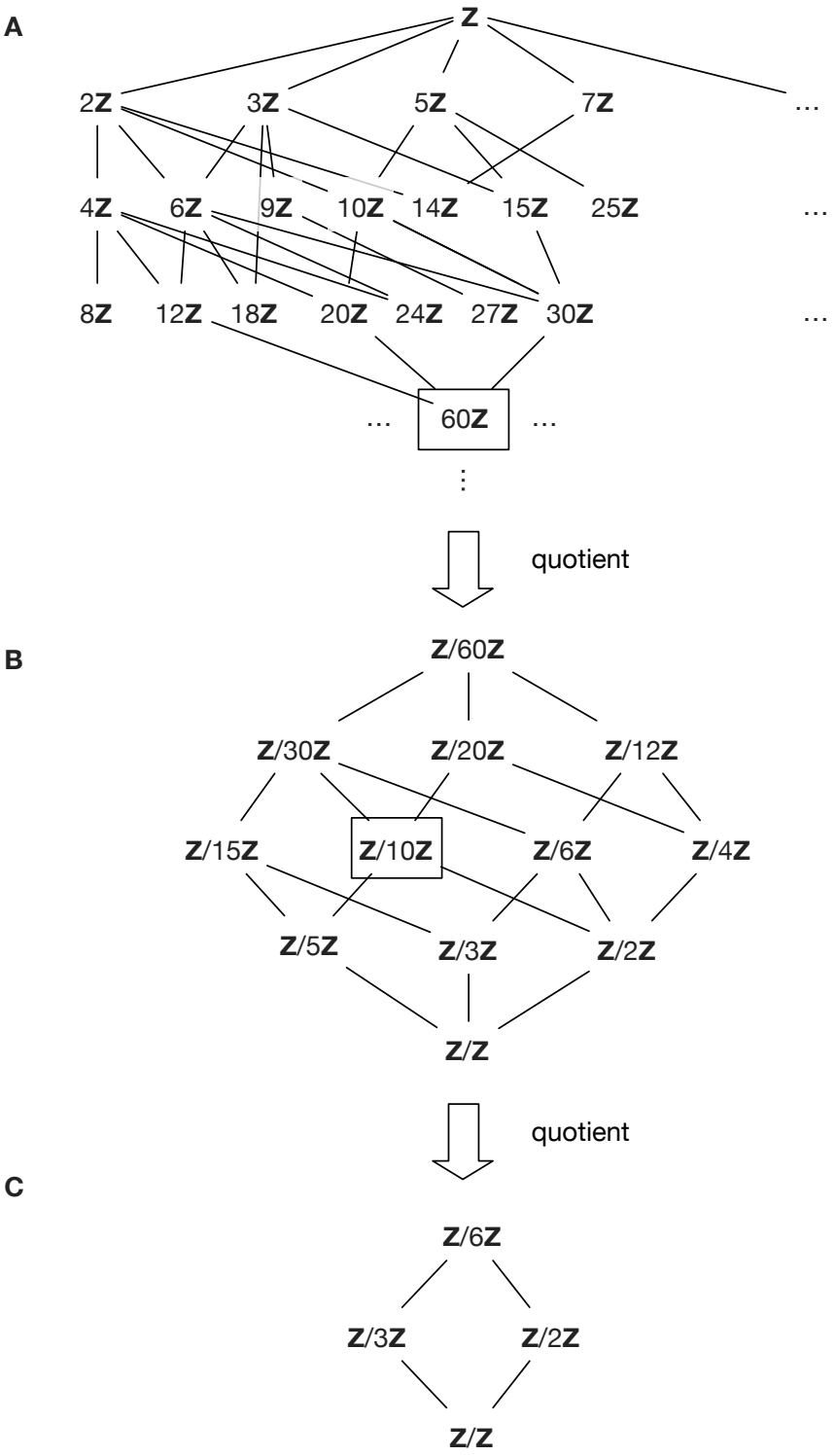


Figure 5.2: Lattice of Congruences on **A** $\mathbb{Z}$, **B** $\mathbb{Z}/60\mathbb{Z}$ and **C** $\mathbb{Z}/60\mathbb{Z}$. The lattice **C** results from **B** by taking the quotients modulo $\mathbb{Z}/10\mathbb{Z}$, which is obtained from **A** as a quotient modulo $60\mathbb{Z}$.

Definition 5.1.11. A *relation* $\mathscr{R}$ on a set S is a subset of the cartesian product

$$\mathscr{R} \subset S \times S.$$

For a pair $(x, y) \in \mathscr{R}$, one writes $x\mathscr{R}y$.

Definition 5.1.12. A *semigroup* is a set S endowed with a binary associative operation $\circ : S \times S \to S$, i.e. $a \circ (b \circ c) = (a \circ b) \circ c$ for all $a, b, c \in S$. $\circ$ is called the multiplication on S.

Definition 5.1.13. A *subsemigroup* S' of a semigroup $(S, \circ)$ is a subset $S' \subset S$ closed under the semigroup operation $\circ$. One writes $S' < S$

Definition 5.1.14. An *equivalence relation* $\mathscr{R}$ on a set S is a relation that is reflexive, symmetric and transitive, i.e.

$$\begin{aligned}(x, x) &\in \mathscr{R} \text{ for all } x \in S\\ (x, y) \in \mathscr{R} &\Rightarrow (y, x) \in \mathscr{R}\\ (x, y), (y, z) \in \mathscr{R} &\Rightarrow (x, z) \in \mathscr{R}.\end{aligned}$$

Remark 5.1.15. Giving an equivalence relation $\mathscr{R}$ is the same as giving a partition of S into disjoint sets $(S_i)_{i\in I}$

$$S = \coprod_{i\in I} S_i.$$

The $(S_i)_{i\in I}$ are called equivalence classes or cosets. Each $x \in S$ is contained in exactly one coset S_i, which then contains all elements $y \in S$ that are related to x, i.e. $y\mathscr{R}x$, and only those. The coset containing x will be denoted as $x\mathscr{R}$. Vice versa, for a given partition $S = \coprod_{i\in I} S_i$, the relation $x\mathscr{R}y \Leftrightarrow \exists i \in I$ such that $x, y \in S_i$ is an equivalence relation.

Definition 5.1.16. A *congruence* $\mathscr{R}$ on a semigroup S is an equivalence relation that is compatible with the semigroup operation, i.e.

$$x\mathscr{R}x' \text{ and } y\mathscr{R}y' \Rightarrow (xy)\mathscr{R}(x'y') \tag{5.1.1}$$

for all $x, x', y, y' \in S$.

Congruences are partially ordered by inclusion as sets (they are subsets of $S \times S$ by definition 5.1.11). Moreover, they form a lattice. An ad hoc definition of lattice is

Definition 5.1.17. A *lattice* $(L, \vee, \wedge)$ is a partially ordered set $(L, \leq)$ such that any two elements $a, b \in L$ have a smallest upper bound u, i.e. $a \leq u$ and $b \leq u$ and u is minimal and unique with this property and a largest lower bound l with the analogous properties. u is called the *join* of a and b and denoted as $a \vee b$ and l is the *meet* and is denoted as $a \wedge b$.

Remark 5.1.18. Let $\mathscr{R}_1$ and $\mathscr{R}_2$ be two congruences. The lattice of congruences has a maximal element $\mathbb{1} = S \times S$ and a minimal element $\Delta = \{(s, s), s \in S\} \subset S \times S$. Thus, the join $\mathscr{R}_1 \vee \mathscr{R}_2$ can be obtained as the intersection of all congruences containing both $\mathscr{R}_1$ and $\mathscr{R}_2$ and the meet $\mathscr{R}_1 \wedge \mathscr{R}_2$ as the union of all congruences contained in $\mathscr{R}_1$ and $\mathscr{R}_2$.

Definition 5.1.19. Let $\mathscr{R}$ be a congruence on a semigroup S. The *quotient semigroup* $S/\mathscr{R}$ is the set of cosets $\{x\mathscr{R}, x \in S\}$ with the operation inherited from S

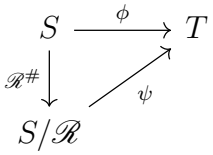
$$(x\mathscr{R})(y\mathscr{R}) = (xy)\mathscr{R}.$$

This operation is well-defined as a consequence of the property 5.1.1 in definition 5.1.16. There is a natural projection from S onto its quotient semigroup

$$\begin{aligned} \mathscr{R}^{\#} : S &\to S/\mathscr{R} \\ x &\mapsto x\mathscr{R}. \end{aligned}$$

Congruences are characterized by the following universal property

Theorem 5.1.20 ([192], Thm. I.5.4)**.** *Let $\mathscr{R}$ be a congruence on a semigroup S. For any semigroup T and homomorphism $\phi : S \to T$ such that $x\mathscr{R}y \Rightarrow \phi(x) = \phi(y)$ there is a unique homomorphism $\psi : S/\mathscr{R} \to T$ such that the diagram*



commutes.

Remark 5.1.21. Let S be a semigroup and $\mathscr{R}$ a congruence. It follows from the previous theorem that there is a one-to-one correspondence between the congruences $\mathscr{R}'$ of S containing $\mathscr{R}$ and the congruences of $S/\mathscr{R}$:

$$\{\mathscr{R}' \text{ such that } \mathscr{R} \subset \mathscr{R}' \subset S \times S\} \leftrightarrow \{\mathscr{R}'/\mathscr{R} \subset S/\mathscr{R} \times S/\mathscr{R}\}$$

by defining $x\mathscr{R}'y$ if and only if $(x\mathscr{R})\mathscr{R}'/\mathscr{R}(y\mathscr{R})$.

Example 5.1.22. As an illustration, example 5.1.10 can be restated in the language of congruences. Defining the congruence $\mathscr{R}_n$ on $\mathbb{Z}$ via

$$a\mathscr{R}_n b \Leftrightarrow a - b \in n\mathbb{Z}$$

identifies the quotient of groups $\mathbb{Z}/n\mathbb{Z}$ with the quotient $\mathbb{Z}/\mathscr{R}_n$. The congruences $\mathscr{R}_n$ form a lattice, whereby the join $\mathscr{R}_m \vee \mathscr{R}_n$ is $\mathscr{R}_{gcd(m,n)}$ and the meet $\mathscr{R}_m \wedge \mathscr{R}_n$ is given by $\mathscr{R}_{smc(m,n)}$ ($gcd(m,n)$ is the greatest common divisor and $smc(m,n)$ is the smallest common multiple of m and n). The lattice of subgroups shown in figure 5.2A corresponds to the lattice of the congruences $\mathscr{R}_n$.

Figure 5.2B shows the lattice obtained from A after taking the join with $\mathscr{R}_{60}$ and then taking the quotients of $\mathbb{Z}$. By remark 5.1.21, the lattice of congruences of $\mathbb{Z}/\mathscr{R}_{60}$ consists of all congruences $\mathscr{R}_n$ containing $\mathscr{R}_{60}$. Taking the quotient of $\mathbb{Z}/\mathscr{R}_{60}$ by $\mathscr{R}_{10}/\mathscr{R}_{60}$ gives the quotient $\mathbb{Z}/\mathscr{R}_6$ leaves the lattice of congruences shown in figure 5.2C.

This example is meant to illustrate that the lattice of congruences of a semigroup contains all possible congruence relations, i.e. quotients that are compatible with the

semigroup operation. After taking a quotient by any given congruence, the congruences of the semigroup larger than the chosen one remain as congruences of the quotient and allow to repeat the procedure.

It is useful to point out that the language of congruences in unnecessary for groups, but is crucial in semigroup theory: For any group G, a congruence $\mathscr{R}$ is uniquely determined by a normal subgroup $N < G$ via $a\mathscr{R}b \Leftrightarrow ab^{-1} \in N$ and each normal subgroup uniquely corresponds to a congruence as the kernel of the projection $G \to G/\mathscr{R}$. For groups, the study of congruences is reduced to the study of normal subgroups. However, for semigroups, it is not the case that congruences are determined by subsemigroups or ideals (although each ideal determines a congruence). For example, congruences on finite semigroups can yield congruence classes of different sizes. This is the case for all Rees quotients of a finite semigroup S by a proper ideal $I \subset S$. Hereby, all elements of $S \setminus I$ form separate classes, whereas all elements of I belong to the same class. In contrast, in quotients of groups G/N all congruence classes are in bijection with the respective normal subgroup N and thus necessarily have the same size.

Remark 5.1.23 (**Biological Motivation).** The preceding example suggests that algebraic structures might be helpful for the coarse-graining of models of biological systems. If a system can be modeled by an algebraic structure such as a semigroup, then the lattice of congruences automatically suggests natural possibilities of coarse-graining.

For example, let A be a set of some system components. If their *interactions* can be described as an operation $\circ$ that leads to other system components of A, i.e. if the interaction between $a \in A$ and $b \in A$ produces a product $c = a \circ b$, then A is a set with an algebraic operation and naturally endowed with a lattice of congruences. Fixing any congruence $\mathscr{R}$, leads to a partition of A into congruence classes and allows to talk about the interactions between the classes.

More specifically, A could be the some set of proteins within a cell and $\mathscr{R}$ the equivalence relation dividing the proteins into classes depending on the protein complex they belong to. If the interaction between proteins $\circ$ can be defined in a physically meaningful such that $\mathscr{R}$ is a congruence, one immediately obtains the interaction between the respective protein complexes. Moreover, the lattice of congruences describes the inclusion of smaller protein complexes into larger ones. Conversely, given a set of proteins A and the interaction $\circ$, the possible coarse-graining procedures compatible with the interaction are given by congruences on A. This idea is sketched in figure 5.3 using $4\mathbb{Z} < \mathbb{Z}$ as a purely algebraic analogy (without biological meaning).

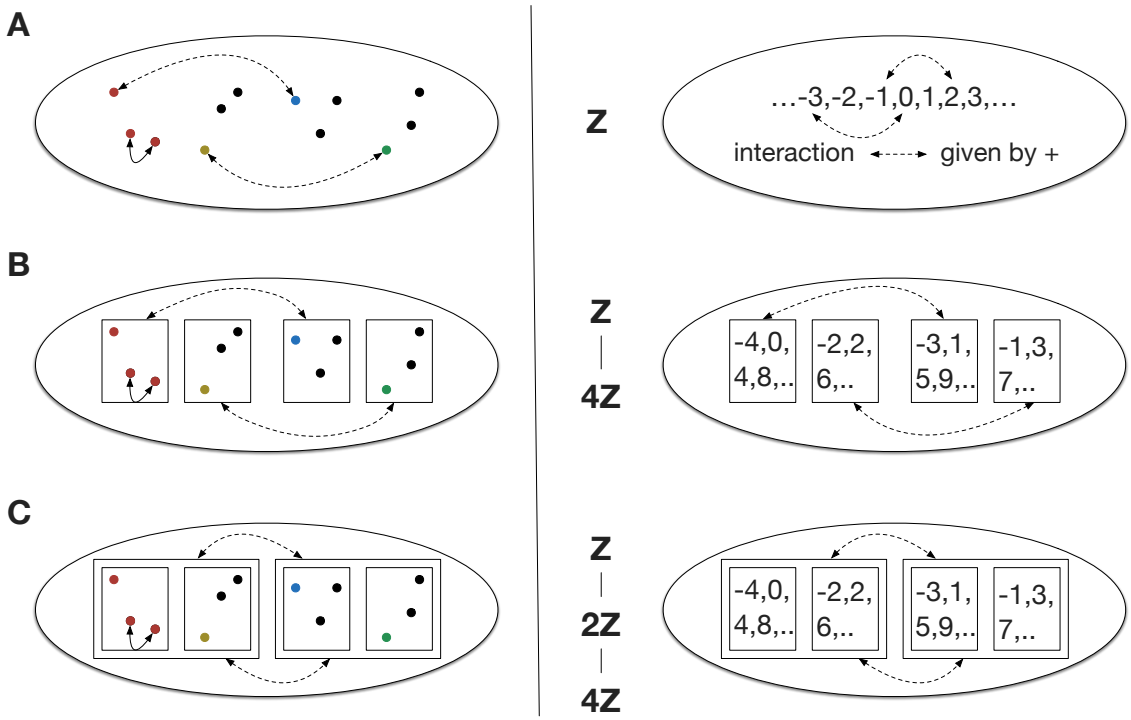


Figure 5.3: Left: **A** Schematic representation of some biological system consisting of a set of components with interactions represented by two-sided arrows. **B** Coarse-graining into four lumped sets with interactions inherited from the component-component interactions. **C** An algebraic procedure automatically suggests further possibilities of coarse-graining and thus shows a hierarchy of nested structures. The interactions within classes as indicated by the solid arrow are only inherited from the algebra if the classes contain idempotents. Right: **A** A far-fetched analogy with the "system" $\mathbb{Z}$ and the "components" $0, \pm 1, \pm 2, ...$ whose interaction is addition. **B** Lumping the elements of $\mathbb{Z}$ into the four residue classes $\bar{0}, \bar{1}, \bar{2}$ and $\bar{3}$ of $\mathbb{Z}/4\mathbb{Z}$. **C** Algebraically, there is an intermediate coarse-graining scheme into $\mathbb{Z}/2\mathbb{Z}$. The addition naturally descends into class $\bar{0}$, but not into the other residue classes.

In general, there is no natural way to assign interactions within the congruence classes. However, congruence classes that contain idempotents (elements e such that $e \circ e = e$) allow the natural interaction within the class inherited from $\circ$: For $a \mathscr{R} e$ and $b \mathscr{R} e$, one obtains $(a \circ b) \mathscr{R} (e \circ e)$ and thus $(a \circ b) \mathscr{R} e$.

The coarse-graining by congruences will applied to semigroup models of CRS in section 5.6.

5.2 Semigroup Models of CRS

The semigroups constructed here combine the formal CRS approach of Kauffman, Hordijk and Steel with the semigroup models constructed by Rhodes [212] and a flavor of Rashevsky's ideas.

The author's main motivation for the use of CRS over classical chemical reaction networks is the possibility to talk about the function of chemical species $x \in X$ and even

about the function of subnetworks and of the whole network on itself.

Throughout this section, let (X, R, C) be a CRS. The state of the CRS is defined by the presence or absence of the chemicals, i.e. by giving the subset $Y \subset X$ of chemicals that are present. Thus the state space $\mathfrak{X}$ of the CRS is the power set $\mathfrak{P}(X) = \{0,1\}^X$. The elements of $\mathfrak{X}$ can be represented by finite tuples $(x_{A_1}, ..., x_{A_n})$ labeled by the set X, i.e. $x_{A_i} \in \{0,1\}$, and $X = \{A_1, ..., A_n\}$. Such tuples $(x_{A_1}, ..., x_{A_n})$ are in one-to-one correspondence with the subsets of X. This correspondence is made explicit by viewing the x_{A_i} in the tuples as the characteristic functions of the singleton sets $\{A_i\}$ giving a bijection

$$\begin{aligned} &\mathfrak{P}(X) \to \{0,1\}^X \\ &\quad Y \mapsto (x_{A_1}(Y), ..., x_{A_n}(Y)) \text{ , where } x_{A_i}(Y) = 1 \text{ iff } A_i \in Y. \end{aligned} \tag{5.2.1}$$

The identification between subsets and tuples will be used interchangeably depending on the context. When reactions $r : X \to \mathbb{Z}$ are directly involved in the construction, the tuple notation is more convenient, but for more abstract constructions and arguments the subset notation is better suited.

A reasonable way to define the function of some given chemical $x \in X$ is via the reactions it catalyzes, i.e. by the way it acts on the state space $\mathfrak{X}$. This definition originates from the work of John Rhodes [212]. The connection to his work is discussed in section 5.7.

Definition 5.2.1. Let (X, R, C) be a CRS with state space $\mathfrak{X} = \{0,1\}^X$. The *function* of $r \in R$ is defined as

$$\phi_r : \mathfrak{X} \to \mathfrak{X}$$

$$\phi_r((x_{A_1}, ..., x_{A_n}))_{A_i} = \begin{cases} 1 & \text{if } A_i \in ran(r) \text{ and } x_{A_j} = 1 \text{ for all } A_j \in dom(r) \\ 0 & \text{else} \end{cases}$$

or, equivalently

$$\phi_r(Y) = \begin{cases} ran(r) & \text{if } dom(r) \subset Y \\ \emptyset & \text{else} \end{cases}$$

for all $Y \subset X$. The sum $\phi + \psi$ of two functions $\phi, \psi : \mathfrak{X} \to \mathfrak{X}$ is defined as

$$(\phi + \psi)((x_{A_1}, ..., x_{A_n}))_{A_i} = \begin{cases} 1 & \text{if } \phi((x_{A_1}, ..., x_{A_n}))_{A_i} = 1 \text{ or } \psi((x_{A_1}, ..., x_{A_n}))_{A_i} = 1 \\ 0 & \text{else,} \end{cases}$$

i.e.

$$(\phi + \psi)(Y) = \phi(Y) \cup \psi(Y)$$

for all $Y \subset X$. The *function* $\phi_x : \mathfrak{X} \to \mathfrak{X}$ of $x \in X$ is defined as the sum over all reactions catalyzed by x

$$\phi_x = \sum_{(x,r)\in C} \phi_r.$$

The functions ϕ_x can be composed via

$$(\phi_x \circ \phi_y)(Y) := \phi_x(\phi_y(Y)) \text{ for any } Y \subset X.$$

This composition $\circ$ is the usual composition of maps and therefore associative. Recalling the definition

Definition 5.2.2 (*)**.** The *full transformation semigroup* $\mathcal{T}(A)$ of a finite discrete set A is the set of all maps $\{f : A \to A\}$ with $\circ$ defined as the composition of maps.

one is led to the definition of the semigroup model for a CRS.

Definition 5.2.3. Let (X, R, C) be a CRS. Its *semigroup model* $\mathcal{S}$ is defined as the semigroup of all maps $\phi : \mathfrak{X} \to \mathfrak{X}$ under composition $\circ$ generated by the $\{\phi_x\}_{x\in X}$ through the operations of composition $\circ$ and union $+$, i.e. $\mathcal{S}$ is the smallest subsemigroup of the full transformation semigroup $\mathcal{T}(\mathfrak{X})$ closed under $\circ$ and $+$ that contains $\{\phi_x\}_{x\in X}$. One writes

$$\mathcal{S} = \langle \phi_x \rangle_{x\in X}$$

As a subsemigroup of $\mathcal{T}(\mathfrak{X})$, $\mathcal{S}$ is automatically a *finite* semigroup.

Remark 5.2.4. By definition, a general map $\phi : \mathfrak{X} \to \mathfrak{X}$ is to be defined on all subsets $Y \subset X$, i.e. the assignment $Y \mapsto \phi(Y)$ needs to be given for all $Y \subset X$. However, in the case of the constructed semigroup models, it is enough to specify the map on some finite set I of generating sets $\{Y_i\}_{i\in I}, Y_i \subset X$ by explicitly defining $\phi(Y_i)$ for all $i \in I$ and by defining

$$\phi(Y) = \bigcup_{Y_i \subset Y} \phi(Y_i).$$

for an arbitrary $Y \subset X$. Usually, the generators $\{Y_i\}_{i\in I}, Y_i \subset X$ will be taken as the sets of substrates of the functions included in ϕ. This is a convenient notational simplification as the state space $\mathfrak{X}$ grows exponentially with the number of chemicals in the network.

Example 5.2.5. As an example, consider the CRS **A** in figure 5.4. Its semigroup model is generated by the maps $\phi_a, \phi_d : \mathfrak{X} \to \mathfrak{X}$. Using the previous remark, the maps will only be specified on their generating sets. The generating set for ϕ_a in the example is $\{c, b\}$ with $\phi_a(\{c, b\}) = \{d\}$. Similarly ϕ_d is generated by $\{a, b\}$ via $\phi_d(\{a, b\}) = \{c\}$. The element $\phi_a + \phi_d$ has both $\{a, b\}$ and $\{c, b\}$ as generating sets with $(\phi_a + \phi_d)(\{c, b\}) = \{d\}$ and $(\phi_a + \phi_d)(\{a, b\}) = \{c\}$. All possible concatenations $\circ$ of any of the maps ϕ_a, ϕ_d and $\phi_a + \phi_d$ yield the zero map $0 : \mathfrak{X} \to \mathfrak{X}$ defined as $0(Y) = \emptyset$ for all $Y \subset X$. This determines the semigroup model $\mathcal{S}$ of the CRS **A** as

$$\mathcal{S} = \{0, \phi_a, \phi_d, \phi_a + \phi_d\} \text{ such that } a \circ b = 0 \text{ for all } a, b \in \mathcal{S}.$$

This is both a left- and right-zero semigroup. The chemical interpretation is that no possible combination of reactions in the CRS produces enough substrates to enable any other reaction within the network. In this particular case, the chemical b is required for all reactions, but is never produced.

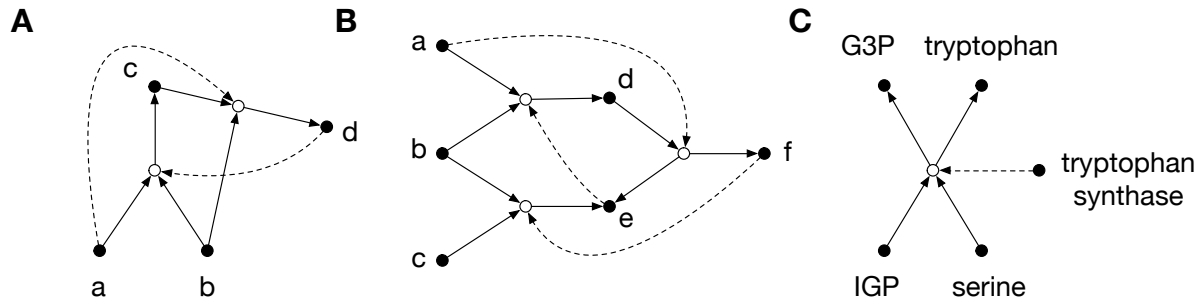


Figure 5.4: Examples of some simple CRS.

The CRS **B** has a nonzero concatenation corresponding to the production of d and e from a, b and c followed by the production of f. In the semigroup language, the map $\phi_a \circ (\phi_e + \phi_f)$ is generated by $\{a, b, c\}$ via $\phi_a \circ (\phi_e + \phi_f)(\{a, b, c\}) = \{f\}$.

One can also recast the chemical reaction network of tryptophan synthase studied in the previous chapter as a CRS. It is shown in figure 5.4**C**. Tryptophan synthase catalyzes the reaction between serine and IGP to form tryptophan and G3P corresponding to the map ϕ_{TS} given by the generator $\phi_{TS}(\{\text{serine, IGP}\}) = \{\text{tryptophan,G3P}\}$. Because $\phi_{TS} \circ \phi_{TS} = 0$ the semigroup model corresponding to the tryptophan synthase Markov network is

$$\mathcal{S} = \{0, \phi_{TS}\} \text{ with } \phi_{TS} \circ \phi_{TS} = 0.$$

The semigroup models in this example are nilpotent semigroups by the

Definition 5.2.6 (*)**.** Let S be a semigroup and n some natural number. The *n-th power* S^n of S is defined as the subsemigroup of S consisting of products of length n

$$S^n = \{a_1 \circ a_2 \circ ... \circ a_n | a_i \in S\}.$$

Definition 5.2.7 (*)**.** A semigroup S is nilpotent if there is an $N \in \mathbb{N}$ such that

$$S^N = \{0\}.$$

For the CRS **A** and **C** in example 5.2.5, one has $\mathcal{S}^2 = \{0\}$ and for **B** one finds $\mathcal{S}^3 = \{0\}$.

The rest of this chapter establishes the basic properties of semigroup models used throughout the chapter. There is natural partial order on $\mathcal{S}$ inherited from the partial order on $\mathcal{T}(\mathfrak{X})$ defined by

Lemma 5.2.8 (*)**.** Let $\phi, \psi \in \mathcal{T}(A)$, where A is a finite and discrete set. Then

$$\phi \leq \psi \Leftrightarrow \phi(B) \subset \psi(B) \text{ for all } B \subset A.$$

is a partial order on $\mathcal{T}(A)$. In particular, this induces a partial order on $\mathcal{S}$. One writes $(\mathcal{S}, \leq)$ for $\mathcal{S}$ endowed with this partial order.

The partial order of a semigroup model $\mathcal{S}$ of a CRS possesses the following property that is not in general valid for transformation semigroups.

Lemma 5.2.9. Let $\mathcal{S}$ be a semigroup model of a CRS. The partial order $(\mathcal{S}, \leq)$ as defined above is preserved under composition, i.e. for any $\phi, \psi, \chi \in \mathcal{S}$

$$\phi \leq \psi \Rightarrow \phi \circ \chi \leq \psi \circ \chi \tag{5.2.2}$$
$$\text{and } \phi \leq \psi \Rightarrow \chi \circ \phi \leq \chi \circ \psi. \tag{5.2.3}$$

Proof. $\phi \leq \psi$ implies $\phi(Y) \subset \psi(Y)$ for all $Y \subset X$ and *a fortiori* $(\phi \circ \chi)(Y) \subset (\psi \circ \chi)(Y)$. This proves 5.2.2. 5.2.3 follows by remark 5.2.4 from $\phi(Y) \subset \psi(Y)$. □

Lemma 5.2.10. Let $\mathcal{S}$ be a semigroup model of a CRS.
(I) Any $\phi, \psi \in \mathcal{S}$ satisfy

$$\phi \leq \phi + \psi. \tag{5.2.4}$$

(II) Any $\phi, \psi, \chi \in \mathcal{S}$ such that $\phi \leq \chi$ and $\psi \leq \chi$ satisfy

$$\phi + \psi \leq \chi. \tag{5.2.5}$$

Proof. This follows directly from remark 5.2.4 and the definition of a sum. □

The operations $\circ$ and $+$ on $\mathcal{S}$ have the following distributivity properties.

Lemma 5.2.11. Let $\phi, \psi, \chi \in \mathcal{S}$. Then

$$\phi \circ \chi + \psi \circ \chi = (\phi + \psi) \circ \chi \tag{5.2.6}$$
$$\text{and } \chi \circ \phi + \chi \circ \psi \leq \chi \circ (\phi + \psi). \tag{5.2.7}$$

Proof. Using the definitions of the operations, one obtains $(\phi \circ \chi + \psi \circ \chi)(Y) = (\phi \circ \chi)(Y) \cup (\psi \circ \chi)(Y) = \phi(\chi(Y)) \cup \psi(\chi(Y)) = (\phi + \psi)(\chi(Y)) = ((\phi + \psi) \circ \chi)(Y)$ for all $Y \subset X$ proving the equality 5.2.6.

Lemma 5.2.9 and lemma 5.2.10(I) imply $\chi \circ \phi \leq \chi \circ (\phi + \psi)$ and $\chi \circ \psi \leq \chi \circ (\phi + \psi)$. 5.2.7 now follows from lemma 5.2.10(II). □

The two operations $\circ$ and $+$ have obvious interpretations in terms of the function of enzymes on a CRS: The sum of two functions $\phi_x + \phi_y$, $x, y \in X$ describes the *joint* or *simultaneous* function of two enzymes x and y on the network - it captures the reactions catalyzed by both x and y at the same time. The sum is associative and commutative by definition. The composition of two functions $\phi_x \circ \phi_y$, $x, y \in X$ describes the *subsequent* function on the network - first y and then x act by their respective catalytic function. Interestingly, using the partial order introduced in lemma 5.2.8, the distributive property 5.2.7 reads: *Applying a test function χ to the sum of two functions ϕ and ψ can be larger*

than applying the test function to the individual functions and then taking the sum. This is reminiscent of the prevalent characterization of emergence (*the whole is larger than the sum of its parts*) and the fact that the simple algebraic models studied here already show this behavior in such clarity is rather surprising to the author.

By definition $\mathcal{S}$ captures all possibilities of joint and subsequent functions of elements of the network on the network itself. In particular, this allows to determine the actions of arbitrary subsets $Y \subset X$ on the whole network by making the

Definition 5.2.12. Let (X, R, C) be a CRS and $Y \subset X$. The semigroup $\mathcal{S}(Y) < \mathcal{S}$ of the functions of Y is

$$\mathcal{S}(Y) = \langle \phi_x \rangle_{x \in Y}$$

and the *function* Φ_Y of Y on X is defined as

$$\Phi_Y = \sum_{\phi \in \mathcal{S}(Y)} \phi.$$

Φ_Y is characterized by the following property.

Proposition 5.2.13. Φ_Y is the unique maximal element of $\mathcal{S}(Y)$ with respect to the partial order introduced in 5.2.8.

Proof. By construction, Φ_Y is an element of $\mathcal{S}(Y)$. It suffices to show that any element $\psi \in \mathcal{S}(Y)$ satisfies $\psi \leq \Phi_Y$. But this is a direct consequence of lemma 5.2.10(I) as $\Phi_Y = \psi + \sum_{\phi \in \mathcal{S}(Y) \setminus \{\psi\}} \phi$ by construction. The unicity follows from the properties of a partial order. □

Remark 5.2.14. In particular, $\mathcal{S}$ has a maximal element Φ_X.

Remark 5.2.15. If $Y \subset Z \subset X$, then the definition 5.2.12 implies $\Phi_Y \leq \Phi_Z$.

One can use the distributivity property 5.2.6 to derive an explicit expression for each $\phi \in \mathcal{S}$ in terms of the functions of chemicals $\{\phi\}_{x \in X}$ as discussed in the following remark. However, this will not be used until section 5.6.

Remark 5.2.16 (Explicit representation of elements of $\mathcal{S}$). Recall that the elements of $\mathcal{S}$ are generated via $+$ and $\circ$ from the functions $\{\phi_x\}_{x \in X}$ of individual chemicals. There is an iterative construction of all elements in $\mathcal{S}$. Denote by $S_0 = \{\phi_x\}_{x \in X}$ and let

$$S_i^\circ = \{ \prod_{\text{finite}} a_j | a_j \in S_{i-1} \} \text{ for } i \geq 1$$

be the set of all possible finite products of elements from S_{i-1}. Let S_i be the set of all possible finite sums of elements from S_i°

$$S_i = \{ \sum_{\text{finite}} a_k | a_k \in S_i^\circ \} \text{ for } i \geq 1.$$

Because $\mathcal{S}$ is a finite semigroup, this construction yields all elements of $\mathcal{S}$ after a finite number of iterations, i.e. there is some $N \in \mathbb{N}$ such that

$$\mathcal{S} = S_N^\circ.$$

Tracing the construction backwards gives the explicit representation of any $\phi \in \mathcal{S} = S_N^\circ$:

$$\phi = a_1 \circ a_2 \circ ... \circ a_n,$$

where all a_j are elements of S_{N-1}, i.e.

$$a_j = \sum_{i_j} a_{ji_j}, a_{ji_j} \in S_{N-1}^\circ.$$

In particular,

$$a_1 = \sum_{i_1} a_{1i_1}, a_{1i_1} \in S_{N-1}^\circ.$$

Distributivity (equation 5.2.6) gives

$$\phi = \sum_{i_1} (a_{1i_1} \circ a_2 \circ ... \circ a_n),$$

where $a_{1i_1} \in S_{N-1}^\circ$. Repeating this at most $N-1$ more times for the leftmost factor gives

$$\phi = \sum_{y \in Y} \phi_y \circ a_2^y \circ ... \circ a_{n_y}^y, \tag{5.2.8}$$

where ϕ_y are functions of single chemicals for some *multiset* Y with elements in X, $a_i^y \in S_{N-1}^\circ$ and the n_y some natural numbers. The same sequence of operations can be repeated for a_2^y giving

$$a_2^y = \sum_{y' \in Y_y} \phi_{y'} \circ a_3^{yy'} \circ ... \circ a_{n_{yy'}}^{yy'}$$

for some multiset Y_y with elements in X, $a_i^{yy'} \in S_{N-1}^\circ$ and the $n_{yy'}$ some natural numbers. Plugging this into the expression 5.2.8 leads to

$$\phi = \sum_{y \in Y} \phi_y \circ (\sum_{y' \in Y_y'} \phi_{y'} \circ a_3^{yy'} \circ ... \circ a_{n_{yy'}}^{yy'}) \circ a_3^y \circ ... \circ a_{n_y}^y.$$

Using distributivity and relabeling the $a_i^y, i \geq 3$ gives

$$\phi = \sum_{y \in Y} \phi_y \circ (\sum_{y' \in Y_y} \phi_{y'} \circ a_3^{yy'} \circ ... \circ a_{m_{yy'}}^{yy'}). \tag{5.2.9}$$

Continuing this resolution for all the remaining functions and taking into account that ϕ was generated by a finite number of operations of taking sums and products (it lies in S_N) implies that ϕ has the form indicated in expression 5.2.9 consisting of consecutive sums of products of functions of single chemicals. This means that ϕ can be represented as a tree with edges labeled by functions ϕ_y and the vertices representing sums over the underlying edges. The sums are then multiplied with the function on the edge above the respective vertex. Figure 5.5A gives an example of such a representation.

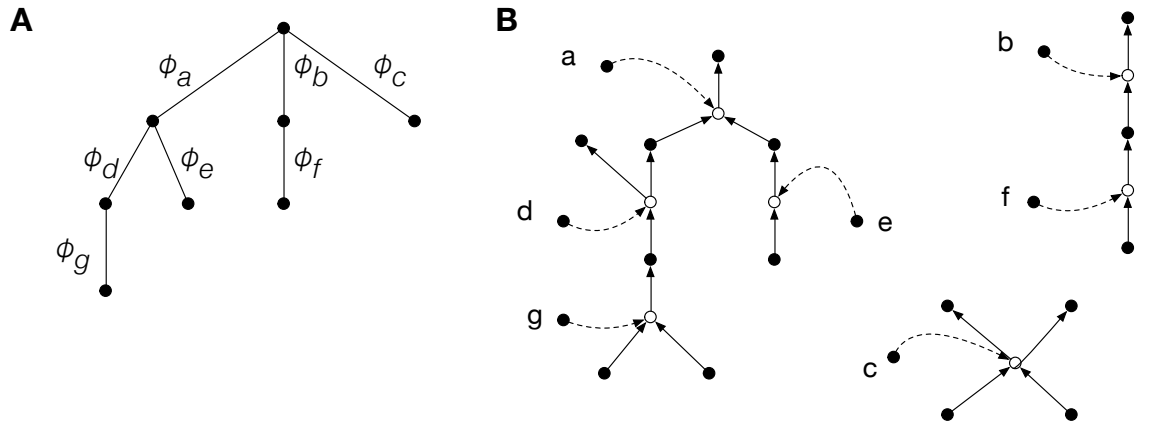


Figure 5.5: The tree **A** shows the function $\phi = \phi_a \circ (\phi_d \circ \phi_g + \phi_e) + \phi_b \circ \phi_f + \phi_c$ (point before line calculation to avoid brackets) as an example of an explicit representation of a general element of $\mathcal{S}$ as discussed in the text. **B** the reaction pathway within a CRS corresponding to the function represented in **A**. As the root of the tree **A** has three branches, the pathway has three components that are not interconnected. Note that the pathway **B** does not represent a unique function. For example, it is also the pathway corresponding to the function $\phi + \phi_g$.

The representation of a function by a tree implies a correspondence to reaction pathways in the CRS, where the leafs of the tree correspond to starting reactions and vertices correspond to joining reaction pathways. As an example, figure 5.5B shows the pathways corresponding to the tree from figure 5.5A. However, the mapping of functions to reaction pathways in neither surjective nor injective in general. In particular, a reaction does not define a unique function. For example, the reaction pathway shown in figure 5.5B corresponds to the function ϕ represented in figure 5.5A, but it is also the reaction pathway of the function $\phi + \phi_g$.

This representation of functions motivates the definition of the support of a given function $\phi \in \mathcal{S}$. Intuitively, this is the minimal set $Y \subset X$ such that ϕ is a function generated by this set. With the notions introduced above, the definition is

Definition 5.2.17. Let ϕ be any function $\phi \in \mathcal{S}$. The *support* of ϕ is the set of minimal sets $Y \subset X$ such that $\phi \in \mathcal{S}(Y)$. It is denoted as $supp(\phi)$.

Remark 5.2.18. The support $supp(\phi)$ can consist of multiple minimal sets that generate ϕ. See figure 5.6 for an example. Therefore, the more appealing definition to require $supp(\phi)$ to be the minimal set $Y \subset X$ such that $\phi \in \mathcal{S}(Y)$ is not well-defined in general. However, if $supp(\phi)$ consists of exactly one set, then the support will be set equal to this set.

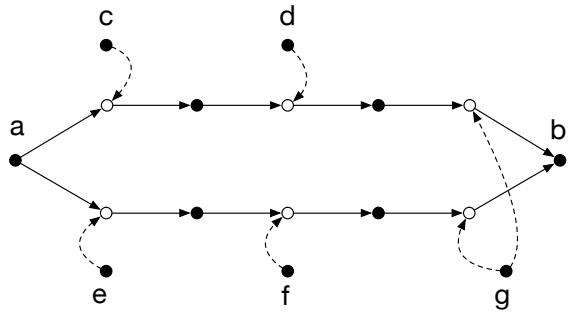


Figure 5.6: An example of a CRS where a function is supported on different minimal subsets of X, i.e. $|supp(\phi)| > 1$. The function ϕ is defined on the generating set $\{a\}$ via $\phi(\{a\}) = \{b\}$. It can be written as $\phi = \phi_g \circ \phi_d \circ \phi_c$ and $\phi = \phi_g \circ \phi_f \circ \phi_e$ and has support $supp(\phi) = \{\{c, d, g\}, \{e, f, g\}\}$.

Note that the support function used here is completely different from the support defined in [189]. In [189], the support is defined on subsets of the reaction set R whereas the support introduced here is defined on elements of $\mathcal{S}$. The definition from [189] is not needed in this work.

5.3 Semigroup Models of CRS with Food Set

Having introduced semigroup models for CRS (X, R, C), the models will now be adapted to CRS with food sets $F \subset X$. The CRS **A** from example 5.2.5 has a semigroup model with $\mathcal{S}^2 = \{0\}$ showing that no combinations of reactions of the network could produce chemicals within the network. Certainly, this CRS is not self-sustaining. However, if the chemicals a and b were constantly supplied, the network would become self-sustaining, i.e. RAF. This should be reflected in an appropriately modified semigroup model.

First, it is not necessary to include the chemicals from the food set in the state space $\mathfrak{X} = \{0, 1\}^X$, because the chemicals from the food set should always be present. Moreover, chemicals that are formed from the food set under reactions catalyzed by the food set need not be included in the state space either, because the also would form in the environment and thus will automatically be externally supplied. This can be achieved by defining the closure of the food set:

Definition 5.3.1. Let (X, R, C, F) be a CRS with food set F. The *closure* $\bar{F}$ is defined as the smallest set containing F such that any reaction r with range outside of $\bar{F}$ requires either a catalyst or a reactant that is not in F.

It is convenient to define the *restriction* of X to F as $X_F := X \setminus \bar{F}$ and the state space $\mathfrak{X}_F := \{0, 1\}^{X_F}$ as the power set of X_F.

Definition 5.3.2. A CRS (X, R, C) with food set F is *RAF* if (X, R, C) is an $\bar{F}$ network according to definition 5.1.5 such that for each element $x \in X_F$ there is a set I of reactions $\{r_i\}_{i \in I}$ producing x and satisfying the conditions (F1) and (F2) from definition 5.1.5 and such that each reaction r_i is catalyzed by some chemical in X.

Remark 5.3.3. If a CRS (X, R, C, F) has a RAF subnetwork, it has a maximal RAF subnetwork as the union of all RAF subnetworks. If its maximal RAF subnetwork is $\bar{F}$, one defines that the CRS (X, R, C, F) has no RAF subnetwork.

To take into account the constant presence of the food set, it is not possible to just replace X by X_F and restrict all maps in $\mathcal{S}$ to X_F, because reactions catalyzed by the food set still need to be included in the model and chemicals in the food set are needed to form chemicals in X_F, yet they do not occur explicitly in X_F. The following definition takes this into account.

Definition 5.3.4. Let (X, R, C) be a CRS with semigroup model $\mathcal{S}$. Let $F \subset X$ be some food set. For each map $\phi \in \mathcal{S}$, the *F-modification* ϕ_F is defined using generating sets introduced in remark 5.2.4. Let $\{Y_i\}_{i\in I}, Y_i \subset X$ be generating sets for ϕ. Then $\{Y_i \cap X_F\}_{i\in I}$ are the generating sets for ϕ_F via

$$\phi_F(Y_i \cap X_F) := \big(\phi(Y_i \cup \bar{F}) \cup \Phi_{\bar{F}}(Y_i \cup \bar{F})\big) \cap X_F,$$

where $\Phi_{\bar{F}}$ is the function of $\bar{F}$ as defined in 5.2.12.

The *semigroup model $\mathcal{S}_F$ of a CRS (X, R, C, F) with food set F* of is a subsemigroup of the transformation semigroup $\mathcal{T}(\mathfrak{X}_F)$ on $\mathfrak{X}_F$ generated by the elements ϕ_F under the operations $+$ and $\circ$, i.e.

$$\mathcal{S}_F = \langle \phi_F \rangle_{\phi\in\mathcal{S}}$$

The semigroup operation is the usual composition $\circ$ inherited from $\mathcal{T}(\mathfrak{X}_F)$.

In the definition of the F-modification ϕ_F of ϕ, the term $\phi(Y_i \cup \bar{F})$ takes into account the constant presence of all elements of $\bar{F}$ as reactants and the term $\Phi_{\bar{F}}(Y_i \cup \bar{F})$ ensures their catalytic action.

Example 5.3.5. As an example for semigroups models with food set, the CRS **A** from example 5.2.5 is reexamined with food set $F = \{a, b\}$ as shown in figure 5.7 and the corresponding semigroup model $\mathcal{S}_F$ with food set is constructed. The maps ϕ_a, ϕ_d and $\phi_a + \phi_d$ have been determined using generating sets in example 5.2.5. Using the definition 5.3.4, the F-modifications ϕ_F are constructed. Afterwards, the closure under $+$ and $\circ$ must be established.

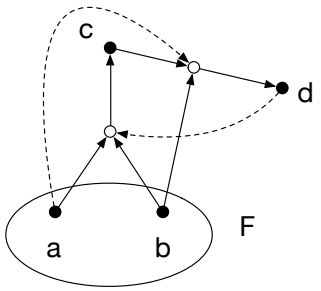


Figure 5.7: CRS **A** from example 5.2.5 with food set $F = \{a, b\}$.

The definition 5.3.1 of the closure of the food set yields $\bar{F} = F$ and $X_F = \{c, d\}$. The F-modifications $(\phi_a)_F, (\phi_d)_F$ and $(\phi_a+\phi_d)_F$ are given by generating sets as $(\phi_a)_F(\{c\}) = \{d\}$, $(\phi_d)_F(\emptyset) = \{c\}$ and $(\phi_a+\phi_d)_F(\emptyset) = \{c\}$; $(\phi_a+\phi_d)_F(\{c\}) = \{c, d\}$. In contrast to example 5.2.5, the concatenations give new elements. It is convenient to introduce a notation for constant maps $c_Y : \mathfrak{X}_F \to \mathfrak{X}_F$ defined by $c_Y(Z) = Y$ for all $Z \subset X_F$ (the zero map 0 is $c_\emptyset$ in this notation). Note that $(\phi_d)_F = c_{\{c\}}$. Some concatenations give more constant elements $(\phi_a + \phi_d)_F^2 = c_{\{c,d\}}$, $(\phi_a)_F \circ (\phi_d)_F = c_{\{d\}}$, $(\phi_a)_F \circ (\phi_a)_F = 0$. The elements $\{0, (\phi_a)_F, (\phi_d)_F, (\phi_a + \phi_d)_F, c_{\{d\}}, c_{\{c,d\}}\}$ are closed under $+$ and $\circ$ as can be seen in the tables 5.1 and 5.2. Thus the semigroup model $\mathcal{S}_F$ is

$$\mathcal{S}_F = (\{0, (\phi_a)_F, (\phi_d)_F, (\phi_a + \phi_d)_F, c_{\{d\}}, c_{\{c,d\}}\}, \circ)$$

with the operation $\circ$ given in table 5.1.

$\circ$	$(\phi_a)_F$	$(\phi_d)_F$	$(\phi_a + \phi_d)_F$	$c_{\{d\}}$	$c_{\{c,d\}}$
$(\phi_a)_F$	0	$c_{\{d\}}$	$c_{\{d\}}$	0	$c_{\{d\}}$
$(\phi_d)_F$	$(\phi_d)_F$	$(\phi_d)_F$	$(\phi_d)_F$	$(\phi_d)_F$	$(\phi_d)_F$
$(\phi_a + \phi_d)_F$	$(\phi_d)_F$	$c_{\{c,d\}}$	$c_{\{c,d\}}$	$c_{\{d\}}$	$c_{\{c,d\}}$
$c_{\{d\}}$	$c_{\{d\}}$	$c_{\{d\}}$	$c_{\{d\}}$	$c_{\{d\}}$	$c_{\{d\}}$
$c_{\{c,d\}}$	$c_{\{c,d\}}$	$c_{\{c,d\}}$	$c_{\{c,d\}}$	$c_{\{c,d\}}$	$c_{\{c,d\}}$

Table 5.1: The multiplication table for $\mathcal{S}_{\{a,b\}}$. The order of composition is *row* $\circ$ *column*.

$+$	$(\phi_a)_F$	$(\phi_d)_F$	$(\phi_a + \phi_d)_F$	$c_{\{d\}}$	$c_{\{c,d\}}$
$(\phi_a)_F$		$(\phi_a + \phi_d)_F$	$(\phi_a + \phi_d)_F$	$c_{\{d\}}$	$c_{\{c,d\}}$
$(\phi_d)_F$			$(\phi_a + \phi_d)_F$	$c_{\{c,d\}}$	$c_{\{c,d\}}$
$(\phi_a + \phi_d)_F$				$c_{\{c,d\}}$	$c_{\{c,d\}}$
$c_{\{d\}}$					$c_{\{c,d\}}$
$c_{\{c,d\}}$					

Table 5.2: The addition table for $\mathcal{S}_{\{a,b\}}$. All functions ϕ satisfy $\phi + \phi = \phi$ giving the corresponding elements on the diagonal. The commutativity of addition yields the lower left half of the table.

For semigroups $\mathcal{S}_F$ of CRS with food set, the lemmata 5.2.8, 5.2.9, 5.2.10 and 5.2.11 remain valid and the analogous proofs hold. Moreover, the definition 5.2.12 of the function $(\Phi_Y)_F$ supported on a subset $Y \subset X$ carries over verbatim and it satisfies the proposition 5.2.13. The representation of a function discussed in remark 5.2.16 and the definition of 5.2.17 with the respective corollaries apply to $\mathcal{S}_F$ as well.

With the construction of a semigroup model $\mathcal{S}_F$ for a CRS with food set, it is possible to give a clean characterization for a CRS to be RAF.

Theorem 5.3.6. *Let (X, R, C, F) be a CRS with food set F and semigroup model $\mathcal{S}_F$. (X, R, C, F) is RAF if and only if Φ_{X_F} is the constant function c_{X_F}, i.e.*

$$\Phi_{X_F}(\emptyset) = X_F.$$

Proof. If (X, R, C, F) is RAF, each chemical is formed by a sequence of catalyzed reactions from $\bar{F}$, i.e. for each $x \in X_F$ there is a function ψ^x such that $x \in \psi^x(\emptyset)$. The function $\Psi := \sum_{x \in X_F} \psi^x$ then satisfies $\Psi(\emptyset) = X_F$. The maximality of Φ_{X_F} yields $\Phi_{X_F} = \Psi$ showing the necessity of the condition.

The condition $\Phi_{X_F}(\emptyset) = X_F$ implies that each chemical in X_F can be formed from $\bar{F}$ by a sequence of reactions catalyzed by elements in X. The representation of Φ_{X_F} as a tree discussed in remark 5.2.16 implies that there is a sequence of reactions satisfying the conditions (F1) and (F2). The partition of the index set required in (F2) is given by the the distance of the function to the root of the tree. □

Corollary 5.3.7. $\Phi_{X_F}(\emptyset)$ contains the maximal RAF.

Proof. Let (X', R', C', F) be the maximal RAF subnetwork with semigroup model $\mathcal{S}'_F$ of the CRS (X, R, C, F) with semigroup model $\mathcal{S}_F$. By definition both CRS have the same food set. As the closure of a food set only depends on the food set, both CRS have the same closure of food sets. Thus $X'_F \subset X_F$ and subsets of X'_F are subsets of X_F and the functions in $\mathcal{S}'_F$ extend to functions on $\mathfrak{X}_F$ as follows: Let ϕ be a function in $\mathcal{S}'_F$, i.e. $\phi : \mathfrak{X}'_F \to \mathfrak{X}'_F$ and define the extension $\phi^e : \mathfrak{X}_F \to \mathfrak{X}_F$ as $\phi^e(Y) = \phi(Y \cap X'_F)$ for all $Y \subset X_F$. In particular, this gives $\Phi^e_{X'_F} \leq \Phi_{X_F}$. Now theorem 5.3.6 implies that $\Phi_{X'_F}(\emptyset) = X'_F$. By construction $\Phi_{X'_F}(\emptyset) = \Phi^e_{X'_F}(\emptyset)$ and therefore $X'_F \subset \Phi_{X_F}(\emptyset)$. □

Corollary 5.3.8. A CRS (X, R, C, F) with nilpotent semigroup $\mathcal{S}_F$ no RAF subnetwork.

Proof. If the CRS had a maximal RAF subnetwork (X', R', C', F), Φ_{X_F} would be bounded from below by the constant function $c_{X'}$ by corollary 5.3.7. Then all powers of Φ_{X_F} would be bounded by $c_{X'}$ as well and therefore $\mathcal{S}_F$ could not be nilpotent. □

5.4 Dynamics on a Semigroup Model

With the tools constructed so far, it is possible to define a discrete dynamics on a CRS with food set by using its semigroup model. The constructions given here are analogously applicable for CRS without a specified food set and will therefore not be mentioned explicitly.

Let (X, R, C, F) be a CRS with food set F and semigroup model $\mathcal{S}_F$. This is the setup for the rest of this section. Starting with any set of chemicals $Y_0 \subset X_F$, there is a maximal function Φ_{Y_0} (definition 5.2.12) that is supported on this set. This function acts on Y_0 giving the maximal set $Y_1 = \Phi_{Y_0}(Y_0)$ that can be produced from Y_0 by using functionality supported only on Y_0 and the food set. The same argument applies to Y_1 and leads to the

Definition 5.4.1. The *discrete dynamics* on a CRS (X, R, C, F) with food set $F \subset X$ with initial condition Y_0 is generated by the propagator $\mathcal{D}$

$$\begin{aligned} \mathcal{D} : \mathfrak{X}_F &\to \mathfrak{X}_F \\ Y &\mapsto \Phi_Y(Y), \end{aligned} \tag{5.4.1}$$

where Φ_Y is the function of $Y \subset X_F$. Analogously, the dynamics is parametrized by $\mathbb{N}$ as

$$Y_{n+1} = \Phi_{Y_n}(Y_n) \text{ for all } n \in \mathbb{N}.$$

Note that the propagator 5.4.1 deletes all elements that are in Y, but not in $\Phi_Y(Y)$.

Remark 5.4.2. Because the state space $\mathfrak{X}_F$ is finite, for the sequence $(Y_n)_{n\in\mathbb{N}}$ there exist minimal natural numbers k and m such that $Y_k = Y_{k+m}$. Taking into account that the dynamics generated by $\mathcal{D}$ is memoryless gives rise to periodic behavior, i.e. $Y_{k+i} = Y_{k+i+nm}$ for all $i = 0, ..., m-1$ and all $n \in \mathbb{N}$. If $m = 1$, Y_k is a fixed point and one says that the dynamics *stabilizes* at Y_k. If $m > 1$, one says that the dynamics has period m and is *oscillatory*. Both behaviours are possible in CRS. According to theorem 5.3.6, if (X, R, C, F) is RAF, then X_F is a fixed point for the dynamics with initial condition $Y_0 = X_F$.

Example 5.4.3. Figure 5.8 shows a CRS with $X = X_F = \{a, b, c\}$, $F = \bar{F} = \emptyset$ and the respective reactions shown in the figure. If the initial condition Y_0 is a proper subset of X_F, the dynamics has period 3. For example, the dynamics generated by $Y_0 = \{a\}$ is

$$\{a\} \mapsto \{b\} \mapsto \{c\} \mapsto \{a\} \mapsto ...$$

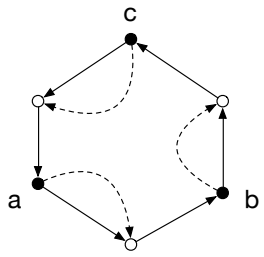


Figure 5.8: Example of a CRS with Possible Oscillatory Dynamics.

ex:period

The discrete dynamics on a CRS can be used to derive further statements about RAF subnetworks of a CRS.

Proposition 5.4.4. Let (X, R, C, F) and let $(Y_n)_{n\in\mathbb{N}}$ be the discrete dynamics with initial condition Y_0. If the semigroup $\mathcal{S}_F$ of the CRS is nilpotent, then the dynamics stabilizes at $\emptyset$, i.e. there exists a natural number N such that

$$Y_n = \emptyset \text{ for all } n \geq N.$$

Proof. By definition $Y_n = \Phi_{Y_{n-1}} \circ \Phi_{Y_{n-2}} \circ ... \circ \Phi_{Y_0}(Y_0)$. Because $\mathcal{S}_F$ is nilpotent, there exists an index N such that $\mathcal{S}_F^N = \{0\}$. This implies that $Y_n = \emptyset$ for all $n \geq N$. □

A useful result is that the dynamics with initial condition X_F cannot have periodic behavior, but always has a fixed point. It is a consequence of the following stronger result.

Proposition 5.4.5. Let (X, R, C, F) be a CRS with dynamics $(Y_n)_{n\in\mathbb{N}}$. If $Y_1 \subset Y_0$, the dynamics is monotonically decreasing, i.e.

$$Y_{n+1} \subset Y_n \text{ for all } n \geq N.$$

Proof. The proof proceeds by induction. By hypothesis $Y_1 \subset Y_0$ is satisfied. Let $Y_n \subset Y_{n-1}$. This implies the ordering of the respective functions $\Phi_{Y_n} \leq \Phi_{Y_{n-1}}$ by remark 5.2.15. This ordering and remark 5.2.4 give the inclusions

$$Y_{n+1} = \Phi_{Y_n}(Y_n) \subset \Phi_{Y_{n-1}}(Y_n) \subset \Phi_{Y_{n-1}}(Y_{n-1}) = Y_n,$$

completing the proof. □

Corollary 5.4.6. With the hypothesis of the previous proposition, the dynamics $(Y_n)_{n\in\mathbb{N}}$ stabilizes.

Proof. By the previous proposition, the dynamics is a descending chain of sets $Y_0 \supset Y_1 \supset \ldots \supset Y_n \supset Y_{n+1} \ldots$. Because X_F is finite, the chain stabilizes. □

Corollary 5.4.7. A dynamics $(Y_n)_{n\in\mathbb{N}}$ with initial condition $Y_0 = X_F$ always leads to a fixed point.

Proof. This follows from $\Phi_{X_F}(X_F) \subset X_F$ and the previous corollary. □

It is convenient to denote the fixed point of the dynamics with initial condition $Y_0 = X_F$ as X_F^* and to refer to X_F^* as *the fixed point* of the CRS. If the CRS is RAF, then $X_F^* = X_F$ by theorem 5.3.6. Intuitively it is clear that X_F^* contains the maximal RAF set of the CRS, because any RAF will constantly reproduce itself. This is made precise in the

Proposition 5.4.8. The fixed point X_F^* of a CRS (X, R, C, F) contains the maximal RAF set.

Proof. Let $(Y_n)_{n\in\mathbb{N}}$ be the discrete dynamics with $Y_0 = X_F$ and let (X', R', C', F) be the maximal RAF subset of (X, R, C, F). If $Y \subset X_F$ contains X'_F, then $\Phi_{X'_F} \leq \Phi_Y$ by remark 5.2.15. In particular, $\Phi_{X'_F}(\emptyset) \subset \Phi_Y(\emptyset) \subset \Phi_Y(Y)$. By theorem 5.3.6 $X'_F \subset \Phi_{X'_F}(\emptyset)$ and thus $X'_F \subset \Phi_Y(Y)$. X'_F is contained in $Y_0 = X_F$ and it follows inductively that $X'_F \subset Y_n$ for all $n \in \mathbb{N}$. By the previous corollary the dynamics stabilizes and thus $X'_F \subset X_F^*$. □

5.5 Identification of RAF Subnetworks

This section uses and compares the tools from the two previous sections to determine the maximal RAF subnetwork of any given CRS. The identification of RAF sets is important in its own right and the approach taken in this work is to establish a correspondence between RAF networks and their respective semigroups. For example, corollary 5.3.8 shows that a CRS with nilpotent semigroup cannot contain any RAF subnetworks. This is an important fact by itself as most semigroups are nilpotent and this weeds out these objects in the study of RAF networks. If the converse were true, it would be possible to tackle the combinatorial properties of RAF sets using the knowledge and tools from semigroup theory, where combinatorial problems are an important and developed field [213]. However, as discussed in the end of this section, the converse is not true within the setup constructed in this work, but many pathological cases can be excluded by thermodynamical considerations.

As in the previous section, let (X, R, C, F) be a CRS with food set and semigroup model $\mathcal{S}_F$. In corollary 5.3.7 and lemma 5.4.8, it was established that both $\Phi_{X_F}(\emptyset)$ and X_F^* contain the maximal RAF subset. As shown in the following example, there is no general relation between the two sets and the inclusion of the maximal RAF subset can be strict. However, as shown in theorem 5.5.5, a combination of the methods used in 5.3.7 and 5.4.8 yields the maximal RAF subset of the CRS.

Example 5.5.1. Figure 5.9 shows two networks with $\Phi_{X_F}(\emptyset)$ and X_F^* given in table 5.3. This shows that the containment of the maximal RAF set is not necessarily strict and that in general neither of $\Phi_{X_F}(\emptyset)$ and X_F^* is contained in the other.

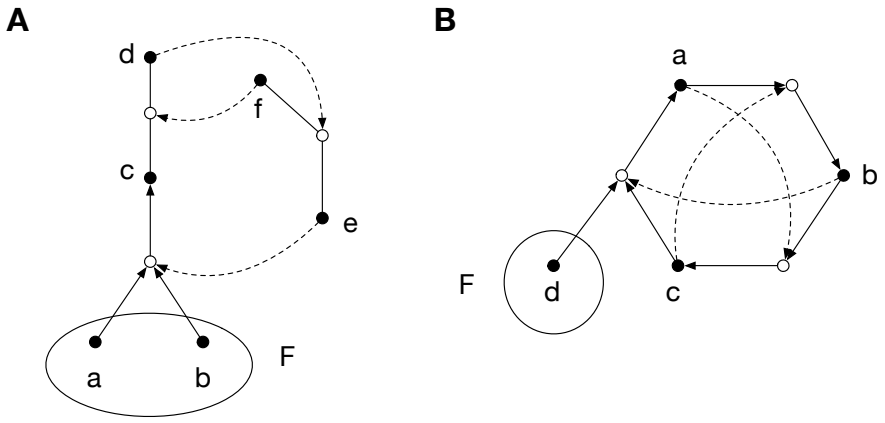


Figure 5.9: Two CRS with food sets demonstrating that there is no relationship between $\Phi_{X_F}(\emptyset)$ and X_F^*.

Network	$\Phi_{X_F}(\emptyset)$	X_F^*	maximal RAF set
A	$\{c, d\}$	$\emptyset$	$\emptyset$
B	$\emptyset$	$\{a, b, c\}$	$\emptyset$

Table 5.3: $\Phi_{X_F}(\emptyset)$, X_F^* and the maximal RAF subset of the networks from figure 5.9.

The CRS 5.9A has two branches $\{c, d\}$ and $\{e, f\}$ that are not connected by chemical reactions. Only the $\{c, d\}$ branch is connected to the food set and all reactions in this branch are catalyzed by Φ_{X_F}. Therefore, $\Phi_{X_F}(\emptyset) = \{c, d\}$. The discrete dynamics starting with the full set $\{c, d, e, f\}$ leads to a depletion of e, then of f and c and then of d giving the empty set as the fixed point.

The system 5.9B has a cyclic arrangement that is self-sustaining and as such $X_F^* = X_F = \{a, b, c\}$. However, none of the chemicals forms from the food set alone and therefore the network has no F network leading to $\Phi_{X_F}(\emptyset) = \emptyset$.

This example shows the essence of the failure for $\Phi_{X_F}(\emptyset)$ and X_F^* to be the maximal RAF subsets. X_F^* contains self-sustaining cycles that contain chemicals not formed from the food set alone and as such do not match the definition of an F set. Φ_{X_F} contains functions that are provided by chemicals not formed from the food set. A combination of the two examples where the $\{e, f\}$ branch of network A is replaced by network B would provide a CRS where both $\Phi_{X_F}(\emptyset)$ and X_F^* are strictly larger than the maximal RAF subset.

The maximal F network of the CRS 5.9B is $\emptyset$ and this implies $\Phi_{X_F}(\emptyset) = \emptyset$ by the following lemma.

Lemma 5.5.2. $\Phi_{X_F}(\emptyset)$ is contained in the maximal F network, i.e. it is an F subnetwork of the CRS.

Proof. Each chemical in $\Phi_{X_F}(\emptyset)$ is formed solely from chemicals in $\bar{F}$. These are by definition formed solely from F. □

Moreover, in the CRS 5.9B, the obstruction for X_F^* to be equal to the maximal RAF set is that the discrete dynamics has initial condition Y_0 not contained in the maximal F network as is implied by the following lemma.

Lemma 5.5.3. Let (X, R, C, F) be a CRS with discrete dynamics $(Y_n)_{n\in\mathbb{N}}$ with a fixed point Y^* such that Y_0 is contained in the maximal F network. Then Y^* is contained in the maximal RAF network.

Proof. Because Y_0 is contained in the maximal F network, one sees inductively that Y^* is contained in the maximal F network. $Y^* = \Phi_{Y^*}(Y^*)$ implies that all reactants are formed from F by some sequence of catalyzed reactions and thus Y^* is RA. □

Remark 5.5.4. If the dynamics $(Y_n)_{n\in\mathbb{N}}$ is periodic, the proposition still applies with an analogous proof.

The main theorem on the maximal RAF subset now follows from the previous results.

Theorem 5.5.5 (on the maximal RAF subset)**.** *For any CRS (X, R, C, F), the maximal RAF subset is the fixed point Y^* of the dynamics $(Y_n)_{n\in\mathbb{N}}$ with initial condition $Y_0 = \Phi_{X_F}(\emptyset)$.*

Proof. First note that

$$Y_1 = \Phi_{Y_0}(Y_0) \subset \Phi_{X_F}(Y_0) = \Phi_{X_F}^2(\emptyset) \subset \Phi_{X_F}(\emptyset) = Y_0,$$

where the containments follow from the maximality of Φ_{X_F}. By proposition 5.4.5, the dynamics has a fixed point Y^*.

By lemma 5.5.2, Y_0 is contained in the maximal F network and thus Y^* is contained in the maximal RAF network by lemma 5.5.3.

$\Phi_{X_F}(\emptyset)$ contains the maximal RAF network by corollary 5.3.7. By the same argument as in the proof of proposition 5.4.8 all Y_n contain the maximal RAF network and so does Y^*. This shows the reverse inclusion.

□

This theorem concludes the formal treatment of the application of semigroup models to RAF sets. In connection to the CRS B in example 5.5.1, it is tempting to discuss the connection between CRS, their maximal RAF subnetworks and the role of thermodynamics in a concluding remark.

Remark 5.5.6 (**Thermodynamics of CRS**)**.** To the author's knowledge, a connection between CRS and classical chemical reaction networks (CRN) has not been established in the literature so far. Viewing CRS as CRN has the advantage of being able to apply

the theory of non-equilibrium thermodynamics for CRN [214] and thus to sort out corresponding CRS that are thermodynamically impossible. The transformation of a CRS into a CRN is rather straightforward. The main idea is to write out the catalytic function of each chemical as a reaction cycle as illustrated in figure 5.10A. Reactions that involve multiple reactants or products lead to multiple or to larger cycles (figure 5.10B). The transformation into catalytic cycles is not unique (cf. A and C in figure 5.10), but the thermodynamic properties of the CRN only depend on the sum of all cycle fluxes corresponding to the particular chemical reaction [214].

Moreover, the food set is considered to be the set of chemicals whose potential is maintained constant by chemostats. Assuming that the reaction network is in a steady-state, these potentials determine the net chemical fluxes for the reaction network. In particular, they determine the direction of the respective cycle fluxes corresponding to the catalyzed chemical reactions. Then the direction of the catalyzed reaction corresponds to the sum of all cycle fluxes. Figure 5.11 shows the cycle decomposition of the CRN 5.10B and C. The CRN 5.10B decomposes into three cycles a, b and c with the respective orientations. Both a and b correspond to the reaction $A + B \to C$, whereas c does not correspond to any transformation. Therefore, the flux of the reaction $A + B \to C$ equals the sum of the fluxes of a and b. Analogously, the CRN 5.10C is decomposed into two cycles d and e. d corresponds to the reaction $A \to B$ and c to its reverse. Therefore the flux of the reaction $A \to B$ is difference of the cycle fluxes of d and e.

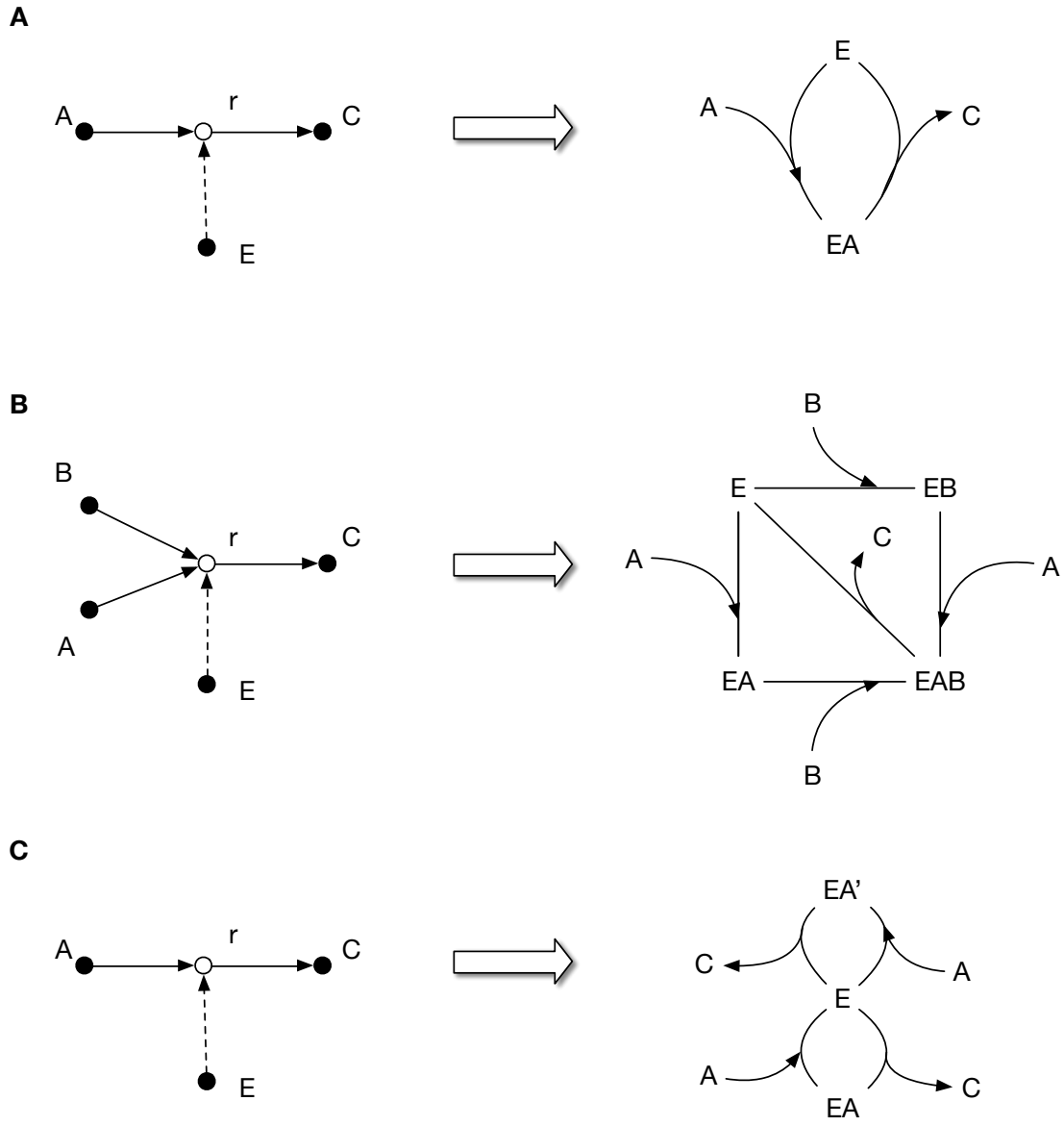


Figure 5.10: Illustration of the conversion of CRS into CRN by expansion of the catalytic cycles. The expansion is not unique as shown by the examples **A** and **C**. However, the thermodynamic properties depend only on the sum of fluxes for all cycles that correspond to a particular reaction.

This setup suggests to think of a CRS without a specified food set as a network of possible chemical reactions with undetermined directionality. The directions of all reactions are only determined upon the choice of food set and the respective chemical potentials. The catalytic cycles without a net flux seem to impose difficulties, because the inclusion of both directed reactions into the CRS could lead to apparent self-sustaining subnetworks that just correspond to catalyzed reactions at equilibrium. However, the values of chemical potentials of the food set that create such situations have measure zero among the space of all possible chemical potentials and therefore it can be safely assumed that such situations do not occur.

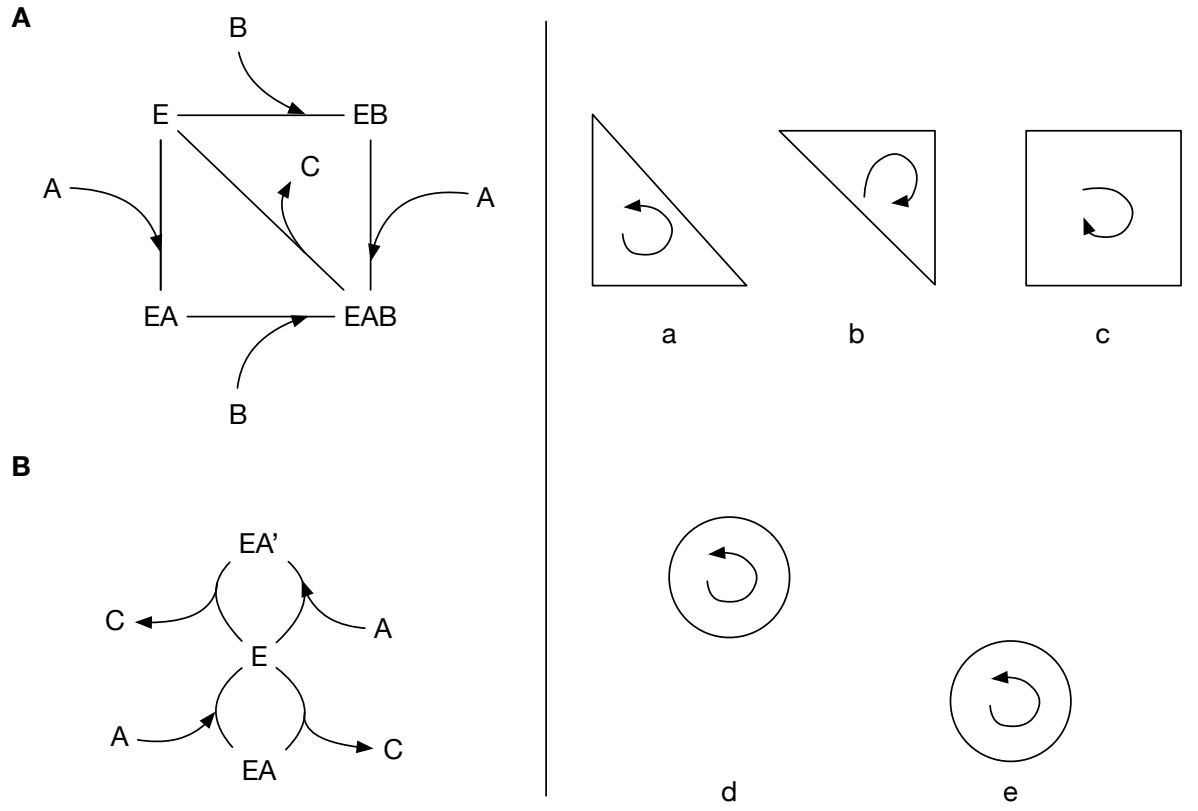


Figure 5.11: Decomposition of CRN into cycles. The net flux of a reaction is determined by the sum of the respective cycle fluxes. The cycle orientations need to be taken into account: Both a and b correspond to the reaction $A + B \to C$ and the flux of the reaction $A + B \to C$ equals the sum of the fluxes of a and b (c corresponds to no transformation). d corresponds to the reaction $A \to B$ and c to its reverse, i.e. the flux of the reaction $A \to B$ is difference of the cycle fluxes of d and e.

From this it follows that self-sustaining cycles that are not linked to chemostats, i.e. to the food set, are thermodynamically impossible. This applies to the example shown in figure 5.8. CRS without such cycles will be called *thermodynamically consistent*. However, the CRS B in example 5.5.1 is thermodynamically consistent and in fact one can construct a CRN corresponding to this CRS. Although the CRS 5.5.1B is not RAF, it is certainly self-sustaining. This applies to all fixed point sets X_F^* of a thermodynamically consistent CRS. Therefore, in future work the author will focus on the class of fixed point sets X_F^* instead of RAF subnetwork as the former capture precisely the notion of self-sustainment whereas the latter are too narrowly defined.

5.6 Algebraic Coarse-Graining

This section revisits the algebraic coarse-graining procedure via congruences sketched in example 5.1.22 and its biological interpretation from remark 5.1.23. It has been illustrated there that congruences of a semigrouop S can be thought of as lumped states or objects

$x\mathscr{R}$ that obey the same operation $\circ$ as the "microscopic" objects $x \in S$.

For general finite semigroups very little is known about the lattice of congruences. Therefore, meaningful results on the structure of congruence lattices of arbitrary $\mathcal{S}_F$ can only be obtained after a deeper understanding of the structure of $\mathcal{S}_F$. As a fist step in this direction, section 5.6.1 gives a proof that all semigroups $\mathcal{S}_F$ of CRS indeed have non-trivial congruences. The given proof heavily relies on the details of the underlying CRS.

In the case semigroups $\mathcal{S}_F$ of CRS, it has to be noted that $\mathcal{S}_F$ does not contain any objects of the CRS, but functions acting on it. In section 5.6.2, examples of biologically meaningful coarse-graining procedures in function are given. Thereby, the support of any function in $\mathcal{S}_F$ allows to relate the function to subsets of chemicals of the network and thus to relate a coarse-graining in function to partitions of the set of chemicals or state of the network. A second family of congruences describes another interesting coarse-graining of the system: The CRS is covered with local patches in a way that the local information on the network is fully retained, while the environment of each patch is no longer resolved.

5.6.1 Existence of Congruences on Semigroup Models

For the rest of the section, let $\mathcal{S}_F$ be a semigroup model of a CRS (X, R, C, F) with food set. Moreover, let the CRS be thermodynamically consistent as discussed in remark 5.5.6. As a first step of the analysis, it is shown that semigroup models of CRS generally have non-trivial congruences. Congruence-free finite semigroups are well-understood. Finite semigroups with 0 have been classified through a structure theorem by Yamura [215] which basically adopts the classification of 0-simple semigroups by Rees [216]. Finite semigroups without 0 admit a neat classification via the theorem

Theorem 5.6.1 ([192], III.6.2.)**.** *A finite semigroup S with $|S| > 2$ either has non-trivial congruences or is a simple group.*

These theorems can be used to show that $\mathcal{S}_F$ has non-trivial congruences in a purely mathematical way. However, the author prefers an argument which directly involves thermodynamic properties of the CRS.

First, some basic definitions are recalled. An ideal of a semigroup is defined via

Definition 5.6.2 (*)**.** Let S be a semigroup. An *ideal* I is a proper subset of S such that

$$SI \cup IS \subset I.$$

where the notation $AB = \{a \circ b | a \in A, b \in B\}$ for $A, B \subset S$ is used. Any ideal $I \subset S$ defines a congruence $\mathscr{R}_I$ as

$$\mathscr{R}_I = \{(x, y) | x, y \in I\} \cup \{(z, z) | z \in S\}, \tag{5.6.1}$$

which is non-trivial if $I \neq \{0\}$. The quotient $S/\mathscr{R}_I$ is also denoted as S/I and called Rees factor semigroup. By finding ideals in $\mathcal{S}_F$, the following theorem can now be proven.

Theorem 5.6.3. *The semigroup $\mathcal{S}_F$ of a thermodynamically consistent CRS (X, R, C, F) with food set admits non-trivial congruences if $|\mathcal{S}_F| > 2$.*

Proof. By the preceding, it suffices to show that $\mathcal{S}_F$ has a nonzero ideal I. Note that $|\mathcal{S}_F| > 2$ is assumed as for $|\mathcal{S}_F| = 1$ and $|\mathcal{S}_F| = 2$, $\mathcal{S}_F$ does not have enough elements to admit non-trivial congruences. 3 cases are considered in the proof.

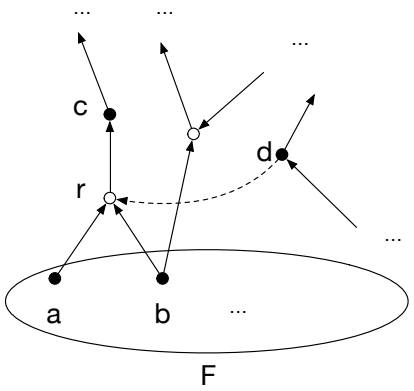


Figure 5.12: Within any CRS, catalyzed reactions r with $dom(r) \subset \bar{F}$ give rise to constant functions. The figure shows a subnetwork of some CRS where this situation occurs. Here ϕ_d is the constant function $c_{\{c\}}$.

Case 1. If $\mathcal{S}_F$ contains nonzero constant functions as well as non-constant functions, the constant functions form an ideal. (Both $\phi \circ c$ and $c \circ \phi$ are constant for a constant function c and any $\phi \in \mathcal{S}_F$.)

Case 2. If $\mathcal{S}_F$ has only constant functions, then any equivalence relation on $\mathcal{S}_F$ is automatically a congruence, because the congruence condition 5.1.1 from definition 5.1.16 is trivially satisfied. As $|\mathcal{S}_F| > 2$, $\mathcal{S}_F$ admits non-trivial equivalence relations.

Case 3. $\mathcal{S}_F$ does not have any constant functions except 0. This case uses details of the CRS underlying $\mathcal{S}_F$. First, any reaction with reactants solely from the food set cannot be catalyzed by any function in $\mathcal{S}_F$. If it was, then there would be some element $x \in X_F$ catalyzing this reaction and its function ϕ_x or some power of it would be nonzero and constant (see figure 5.12 for an illustration).

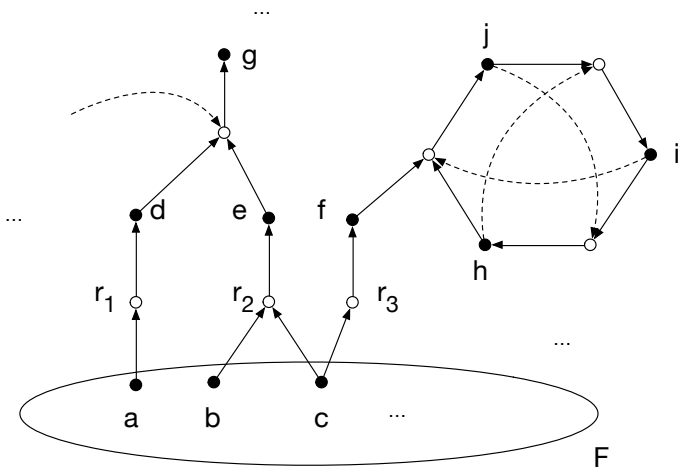


Figure 5.13: If the CRS has no nonzero constant functions and for each reaction r of CRS, either $dom(r) \subset \bar{F}$ or $dom(r) \subset X_F$, then its semigroup $\mathcal{S}_F$ is nilpotent. The figure shows a subnetwork of some CRS. The reactions r_1, r_2 and r_3 all have $dom(r_i) \subset \bar{F}, i = 1, 2, 3$ and therefore cannot be catalyzed. In this subsnetwork $\Phi_{X_F}(X_F)$ does not contain d, e and f, $\Phi^2_{X_F}(X_F)$ does not contain g and j either and $\Phi^4_{X_F} = 0$.

Case 3.1. Assume first that all elements in the food set react only with each other and not with chemicals in X_F, i.e. if each reaction r of CRS has either $dom(r) \subset \bar{F}$ or $dom(r) \subset X_F$. This means that the application of Φ_{X_F} to X_F will deplete X_F by the chemicals formed directly from the food set, an application of Φ_{X_F} to the resulting set will deplete it by all elements formed from $\bar{F}$ by two successive reactions and iteratively $\Phi^N_{X_F}$ will deplete X_F by all chemicals formed from $\bar{F}$ by N successive reactions. By remark 5.5.6, all directed reactions in a thermodynamically consistent CRS must be linked to the food set and therefore there exists an N such that $\Phi^N_{X_F} = 0$ (see figure 5.13 for an illustration). $\Phi^N_{X_F}$ is the maximal element of $\mathcal{S}^N_F$ and thus $\mathcal{S}^N_F = \{0\}$. For a nilpotent semigroup, either $\mathcal{S}^2_F = \{0\}$ or $\mathcal{S}^2_F$ is a proper ideal of $\mathcal{S}_F$. ($\mathcal{S}^2_F$ is an ideal by definition. If it was not proper, then $\mathcal{S}_F = \mathcal{S}^2_F = ... = \mathcal{S}^N_F$ for any N.) If $\mathcal{S}^2_F = \{0\}$, then any non-trivial equivalence relation yields a congruence on $\mathcal{S}_F$, because the congruence condition 5.1.1 from definition 5.1.16 is trivially satisfied. Otherwise the ideal $\mathcal{S}^2_F$ gives a non-trivial congruence.

Case 3.2. Assume now that there are reactions where chemicals from $\bar{F}$ and X_F react with each other and that $\mathcal{S}_F$ is not nilpotent. The case of nilpotent $\mathcal{S}_F$ can be treated as above. This implies that there is a cyclic subnetwork $Y \subset X_F$ that is linked to the food set and all whose reactions are catalyzed by some chemical of X_F. The condition of being cyclic is necessary since no power of Φ_{X_F} is zero and therefore $Y \subset \Phi_{X_F}(Y)$ must be satisfied. This is illustrated in figure 5.14. Choose a minimal Y with this property, i.e. such that for all $y \in Y$ one has $Y \subset \Phi_{X_F}(Y)$, but $Y \setminus \{y\} \not\subset \Phi_{X_F}(Y \setminus \{y\})$. Without loss of generality one can assume that $Y = X_F$. This implies that $\Phi_{X_F}(X_F) = X_F$ and $X_F \setminus \{x\} \not\subset \Phi_{X_F}(X_F \setminus \{x\})$ for all $x \in X_F$. In particular $\Phi_{X_F}(X_F \setminus \{x\})$ is a proper subset of X_F. Now the following elementary lemma gives the desired result.

Lemma 5.6.4 ([192], III.1.3.)**.** A semigroup S is simple (i.e. contains no proper ideals) if and only if $SaS = S$ for all elements $a \in S$. Or equivalently, if and only if for all $a, b \in S$

there exist $x, y \in S$ such that $xay = b$.

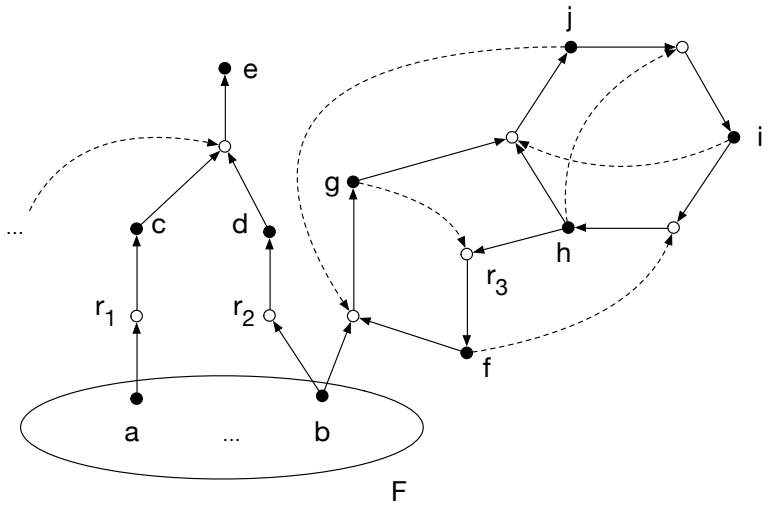


Figure 5.14: A CRS without nonzero constant functions can have a non-nilpotent semigroup model $\mathcal{S}_F$. This implies the presence of a cyclic subnetwork $Y \subset X_F$ such that $Y \subset \Phi_{X_F}(Y)$. In the case of the subnetwork shown, $Y = \{f, g, h, i, j\}$ is such a cyclic network. The functions r_1 and r_2 cannot be catalyzed and therefore the left linear reaction branch will vanish for some power $\Phi^N_{X_F}$. This CRS has no constant functions as $c \leq \Phi_{X_F}$ for any constant function, thus $\Phi_{X_F}(\emptyset) = \emptyset$ implies $c = 0$.

With the notation of lemma 5.6.4, take $b = \Phi_{X_F}$ and a some function ϕ_c of a chemical c that does not form some reactant $x \in X_F$ (in figure 5.14, any chemical will suffice as c). With the notation introduced in section 5.3, the element ϕ_c should be written as $(\phi_c)_F$, but the subscript will be dropped here to avoid notational overload. If $\mathcal{S}_F$ was simple, then one could find $\phi, \psi \in \mathcal{S}_F$ such that $\Phi_{X_F} = \phi\phi_c\psi$. Using the maximality of Φ_{X_F} this gives

$$\Phi_{X_F} = \phi\phi_c\psi \leq \Phi_{X_F}\phi_c\Phi_{X_F} \leq \Phi_{X_F} \Rightarrow \Phi_{X_F}\phi_c\Phi_{X_F} = \Phi_{X_F}.$$

Applying both maps to the set X_F gives $\Phi_{X_F}(X_F) = X_F$ on the right hand side, but $\Phi_{X_F}\phi_c\Phi_{X_F}(X_F) = \Phi_{X_F}(\phi_c(X_F))$ on the left hand side. By the above, ϕ_c does not produce x, i.e. $\phi_c(X_F) \subset X_F \setminus \{x\}$ and therefore $\Phi_{X_F}(\phi_c(X_F)) \subset \Phi_{X_F}(X_F \setminus \{x\})$ is a proper subset of X_F. This shows that $\mathcal{S}_F$ is not simple and completes the proof.

□

5.6.2 Constructions of Congruences

As before, let $\mathcal{S}_F$ be a semigroup model of a CRS (X, R, C, F) with food set. A congruence $\mathscr{B}$ related to the organization of metabolic pathways within the CRS and a family $\mathscr{R}_n$, $n \in \mathbb{N}$ of congruences related to the local structure of the CRS are introduced here.

The Subsemigroup of Constant Functions and Metabolic Pathways

The notion of metabolic pathways naturally arises as a congruence on the subsemigroup of $\mathcal{S}_F$ formed by constant functions. For the sake of simplicity it will be assumed that (X, R, C, F) is a RAF set, but generalizations to arbitrary CRS are straightforward. Let $\mathcal{S}_c < \mathcal{S}_F$ be the subsemigroup of constant functions. It is non-empty as Φ_{X_F} is the constant function c_{X_F} by theorem 5.3.6. As discussed in the proof of theorem 5.6.3, case 2, any equivalence relation of $\mathcal{S}_c$ is already a congruence, i.e. any partition of $\mathcal{S}_c = \coprod S_i$ into subsets $\{S_i\}_{i\in I}$ gives a congruence. Thus the number of congruences grows exponentially with $|\mathcal{S}_c|$ and the difficulty lies in the identification of the biologically interesting ones. As the following considerations show, this can be done using the partial order of functions and their support defined in 5.2.17.

To avoid unnecessary technicalities, assume that the RAF network (X, R, C, F) has all reactions catalyzed by exactly one chemical (condition 1) and that each chemical is formed by a unique path of catalyzed reactions within the network (condition 2). The two conditions are fulfilled for reaction networks encountered in biology. Let $\mathscr{B}$ be a biologically meaningful congruence. First, it is reasonable to impose that $\mathscr{B}$ is compatible with the partial order on $\mathcal{S}_c$ inherited from $\mathcal{S}_F$, which is explicitly given by

$$c_Y \leq c_Z \Leftrightarrow Y \subset Z$$

for constant functions. For each chemical $x \in X_F$, there is a minimal constant function c^x forming that particular chemical by the RAF property. By condition 2 this function is unique and by condition 1 it has a unique support $supp(c^x)$. The set of support sets $\mathcal{M} = \{supp(c^x)\}_{x\in X_F}$ is partially ordered by inclusion. Figure 5.15 shows an example of a subnetwork of a RAF network and the induced partial order on $\mathcal{M}$. In general, $\mathcal{M}$ does not necessarily contain upper or lower bounds for any two elements $Y, Z \in \mathcal{M}$. For a given element $Y \in \mathcal{M}$, a *successor* of Y is an element $Z \in \mathcal{M}$ such that $Y \leq Z$ and there exists no element $T \in \mathcal{M}, T \neq Y, Z$ such that $Y \leq T \leq Z$. A *precursor* of Y is defined analogously. One says that there is a *fork* with multiplicity n at $Y \in \mathcal{M}$ if Y has multiple successors $\{Y_1, Y_2, ..., Y_n\}$ such that Y is the only precursor for each $Y_i, i = 1, ..., n$. If $Y \in \mathcal{M}$ is the successor of multiple precursors $\{Y_1, Y_2, ..., Y_n\}$, then Y is said to be the *hub* with multiplicity n of $\{Y_1, Y_2, ..., Y_n\}$.

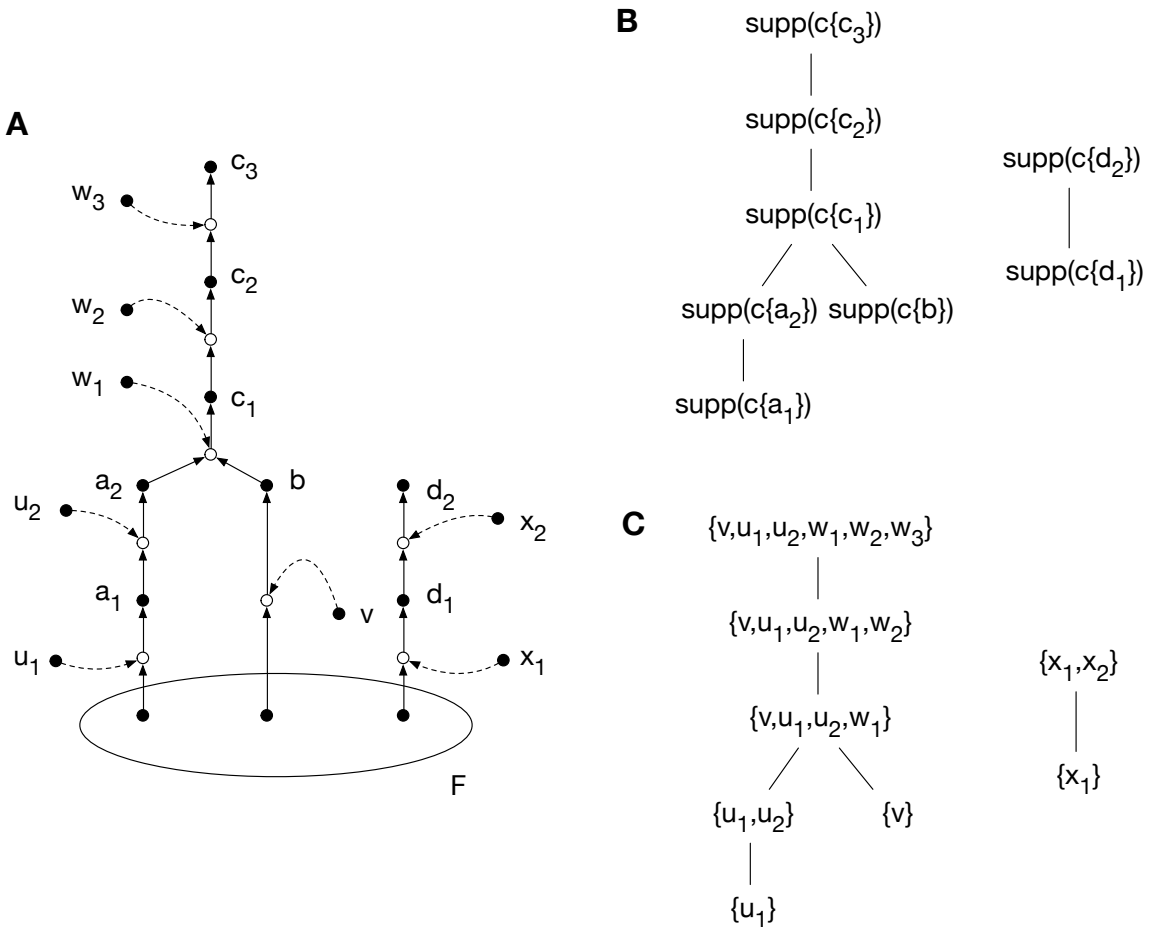


Figure 5.15: **A** Example of metabolic pathways within a RAF set. **B** The set $\mathcal{M}$ corresponding to **A** and its partial order. **C** Explicit representation of the sets from **B**.

The partial order on $\mathcal{M}$ already contains all the information about metabolic pathways and their relations. $supp(c^x)$ is minimal if and only if x is formed by a single reaction with all its substrates contained in $\bar{F}$. Any minimal $supp(c^x)$ with a successor $supp(c^y)$ such that $supp(c^x)$ is the unique precursor of $supp(c^y)$ is part of a linear reaction pathway wherein x is formed from substrates contained in $\bar{F}$ and then further transformed. In such a situation there is a unique chemical z with function ϕ_z such that

$$supp(c^y) = supp(c^x) \cup \{z\}$$

and

$$c^y = \phi_z \circ c^x.$$

Therefore, for all minimal $supp(c^x)$ there are maximal chains of successors corresponding to linear reaction pathways

$$\phi_{z_n} \circ \phi_{z_{n-1}} \circ ... \circ \phi_{z_1} \circ c^x.$$

This allows forks, but no hubs within such linear pathways. Hereby each fork of multiplicity n leads to the branching of one linear pathway into n distinct ones. A hub

$Y = supp(c^y)$ of $\{Y_1, Y_2, ..., Y_n\}, Y_i = supp(c^{x_i})$ corresponds to a reaction with substrates produced in the pathways $Y_1, ..., Y_n$ catalyzed by some chemical z. In this case, one has

$$Y = \{z\} \cup \bigcup_{i=1}^{n} Y_i$$

and

$$c^y = \phi_z \circ \sum_{i=1}^{n} c^{x_i}.$$

This shows that the partial order on the support functions in $\mathcal{M}$ corresponds directly to the organization of connected metabolic pathways within the RAF network. Hubs correspond to reactions that combine products from multiple reaction pathways and forks corresponds to the splitting of a pathway. The structure of reaction pathways that are not connected is not captured by $\mathcal{M}$ (in the example shown in figure 5.15, the support of $c_{\{c_3,d_2\}}$ is not in $\mathcal{M}$). Moreover, the the resolution of linear reaction pathways is too fine. The linear pathways can be contracted by deleting all support sets $Y_1, Y_2, ..., Y_{n-1}$ from $\mathcal{M}$, where $Y_1 \leq Y_2 \leq ... \leq Y_n$ is a linear pathway without forks or hubs. This is achieved through the definition

$$\mathcal{M}' = \mathcal{M} \setminus \left\{ \bigcup_{i=1}^{n-1} Y_i | Y_i \in \mathcal{M} \text{ such that } Y_1 \leq Y_2 \leq ... \leq Y_n \text{ has no forks or hubs} \right\}.$$

To take into account pathways that are not connected, define

$$\mathcal{M}^* = \mathcal{M} \cup \left\{ \bigcup_{i \in I} Y_i | Y_i \in \mathcal{M}' \text{ such that } \forall i, j \in I \text{ there is no } Y \in \mathcal{M}' \text{ such that } Y_i \cup Y_j \subset Y \right\}.$$

The sets $\mathcal{M}'$ and $\mathcal{M}^*$ corresponding to the network from figure 5.15A are shown in figure 5.16. Note that $\mathcal{M}^*$ is a join semilattice in contrast to $\mathcal{M}$ and $\mathcal{M}'$.

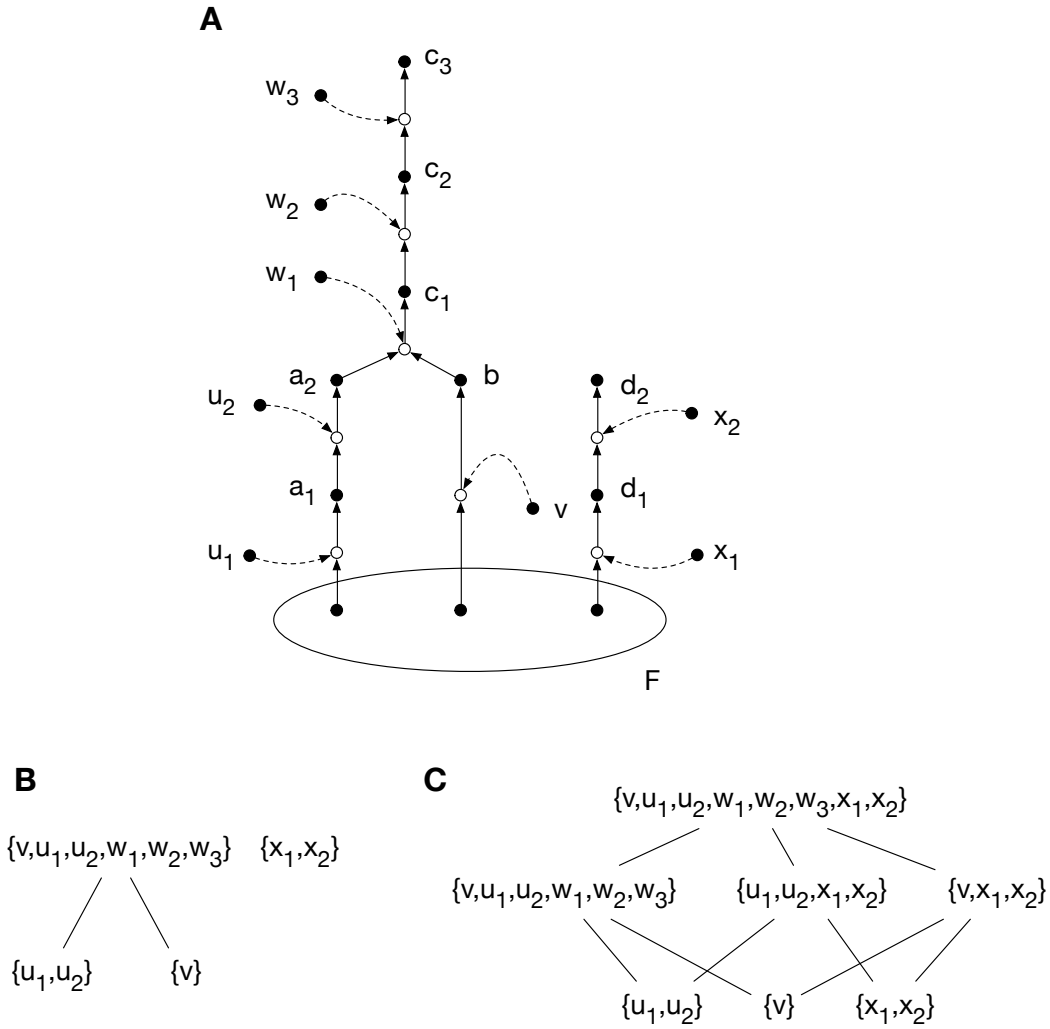


Figure 5.16: **A** The network from figure 5.15. **B** The partially ordered set $\mathcal{M}'$ after contraction of the linear pathways from $\mathcal{M}$. **C** The resulting join semilattice $\mathcal{M}^*$.

At this point, it is straightforward to group the constant functions into equivalence classes depending on the position of their support within $\mathcal{M}^*$ leading to the congruence $\mathscr{B}$.

For all constant functions c_Z, let $Y(Z) \in \mathcal{M}^*$ be the unique minimal element such that $supp(c_Z) \subset Y(Z)$ and define

$$c_Z \mathscr{B} c_{Z'} \Leftrightarrow Y(Z) = Y(Z'). \tag{5.6.2}$$

According to remark 5.2.16, all functions correspond to some set of reaction pathways in the CRS. But for all possible reaction pathways of constant functions there is a minimal element in $\mathcal{M}^*$ that contains their support by construction. Thus the definition 5.6.2 has assigned each element of $\mathcal{S}_c$ to some congruence class of $\mathscr{B}$. The partial order on the functions induces a partial order on the congruence classes of $\mathscr{B}$ via

$$\begin{aligned}(\psi\mathscr{B}) \leq (\psi\mathscr{B}) \Leftrightarrow \text{for all } \phi' \in (\phi\mathscr{B}), \psi' \in (\psi\mathscr{B}) \\ \exists \phi'' \in (\phi\mathscr{B}), \psi'' \in (\psi\mathscr{B}) \text{ such that } \phi' \leq \psi'' \text{ and } \phi'' \leq \psi'\end{aligned}$$

giving a partial order on $\mathcal{S}_c/\mathscr{B}$. This corresponds to the partial order on $\mathcal{M}^*$. The partial order on $\mathcal{S}_c/\mathscr{B}$ describes the hierarchy of possible metabolic pathways and the partial order on $\mathcal{M}^*$ describes the different coarse-graining schemes on X_F that give rise to the respective reaction pathways. In other words, $\mathcal{M}^*$ shows all the subsets of X_F that are *functionally* related and in additions reveals the hierarchy of such relation.

In this example, the semigroup structure is easy to understand from an algebraic point of view: The semigroup of constant functions is a left zero semigroup, i.e. a semigroup S such that $xy = x$ for all $x, y \in S$. Mathematically, there is no reason to prefer some congruence over any other. However, the partial order on the functions and its connection to X_F via the support function gives rise to the biologically interesting congruence $\mathscr{B}$. The semigroup operation descends to an operation on the quotient $\mathcal{S}_c/\mathscr{B}$, but it does not have an interesting biological interpretation and therefore is not discussed further.

Remark 5.6.5. The congruence $\mathscr{B}$ on $\mathcal{S}_c$ cannot be extended to a congruence on $\mathcal{S}_F$, because all congruence classes that contain more than one element would collapse to 0 as the following argument shows. Let $(c_Y\mathscr{B})$ be a congruence class with more than two elements. It contains a constant function $c_{Y'}$ such that Y' is the product set of *one* chemical reaction and $supp(c_{Y'})$ is maximal among the support sets in the congruence class. By definition of $\mathscr{B}$, the class also contains an element $c_{Y'\cup Y''}$, where Y'' is a set of reactants for some $\phi \in \mathcal{S}_F$ such that $supp(\phi) \subset supp(c_{Y'})$ and Y'' does not contain Y'. It follows that $\phi \circ c_{Y'\cup Y''}$ is a constant function contained in the class $(c_Y\mathscr{B})$. This is illustrated in figure 5.17.

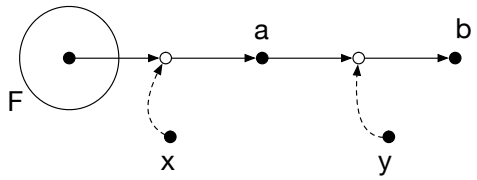


Figure 5.17: With the notations in the text, $Y' = \{b\}$, $Y'' = \{a\}$, $c_{Y'} = \phi_y \circ \phi_x$, $c_{Y'\cup Y''} = \phi_y \circ \phi_x + \phi_x$ and $\phi = \phi_y$ is an example of the general setup.

If there was a congruence $\mathscr{B}'$ on $\mathcal{S}_F$ extending $\mathscr{B}$, then $c_{Y'}\mathscr{B}'c_{Y'\cup Y''}$ would imply

$$\phi \circ c_{Y'}\mathscr{B}'\phi \circ c_{Y'\cup Y''},$$

where the $\phi \circ c_{Y'}$ is the zero function and thus $(c_Y\mathscr{B})$ is the congruence class of 0.

However, imposing that all constant functions $c_Y \in \mathcal{S}_F$ are in the zero congruence class gives rise to the quotient semigroup $\mathcal{S}_F/\mathcal{S}_c$. Congruences on this quotient lead to interesting coarse-graining schemes via the complexity of functions.

Congruences via Complexity of Functions

Let (X, R, C, F) be a RAF network with semigroup model $\mathcal{S}_F$. Due to the finiteness of $\mathcal{S}_F$ the chain $\mathcal{S}_F \supsetneq \mathcal{S}_F^2 \supsetneq ...$ stabilizes for some $N \in \mathbb{N}$, i.e.

$$\mathcal{S}_F \supsetneq \mathcal{S}_F^2 \supsetneq ... \supsetneq \mathcal{S}_F^N = \mathcal{S}_F^{N+1}.$$

As discussed in the proof of theorem 5.6.3, case 3.2, $\mathcal{S}_F^N$ contains all constant functions and non-constant functions corresponding to self-sustaining cycles. Note that the RAF property imposes that $\mathcal{S}_F$ contains none of the latter and therefore $\mathcal{S}_F^N = \mathcal{S}_c$. This suggests the following definition.

Definition 5.6.6. Let ϕ be some function in the semigroup model $\mathcal{S}_F$ of a RAF network. ϕ has *complexity* n if there exists some $n, 1 \leq n \leq N$ such that

$$\phi \in \mathcal{S}_F^n \setminus \mathcal{S}_F^{n+1}.$$

Constant functions (including 0) have complexity ∞. The complexity of ϕ is denoted as $comp(\phi)$.

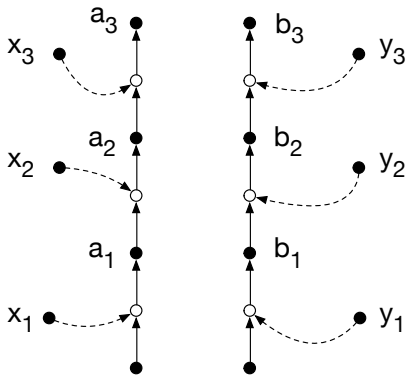


Figure 5.18: The functions $\phi = \phi_{x_2} \circ \phi_{x_1} + \phi_{y_1}$ and $\psi = \phi_{x_3} + \phi_{y_3} \circ \phi_{y_2}$ have complexity 1, but their composition has complexity 3.

The complexity $comp(\phi)$ of a function ϕ determines whether the function can be decomposed into a product of at most $comp(\phi)$ functions. For example, a non-constant function ϕ_x of a chemical $x \in X_F$ has complexity 1 in general as it cannot be further decomposed. By remark 5.2.16, functions correspond to reaction pathways within the CRS. Intuitively, $comp(\phi)$ gives the length of the shortest pathway described by ϕ as illustrated in the

Remark 5.6.7. By definition, any two functions $\phi, \psi \in \mathcal{S}_F$ satisfy $comp(\phi) + comp(\psi) \leq comp(\phi \circ \psi)$. The inequality can be strict as the example in figure 5.18 shows. The functions $\phi = \phi_{x_2} \circ \phi_{x_1} + \phi_{y_1}$ and $\psi = \phi_{x_3} + \phi_{y_3} \circ \phi_{y_2}$ have complexity 1. Their composition can be written as $(\phi_{x_3} + \phi_{y_3}) \circ (\phi_{x_2} + \phi_{y_2}) \circ (\phi_{x_1} + \phi_{y_1})$ and thus has complexity 3.

The powers $\mathcal{S}_F^n$ are proper ideals of $\mathcal{S}_F$ for $2 \leq n \leq N$ and give rise to a congruence $\mathscr{R}_{\mathcal{S}_F^n}$ via the expression 5.6.1. Such congruences will be denoted as $\mathscr{R}_n$ for notational convenience. The resulting quotient semigroups $\mathcal{S}_F/\mathscr{R}_n$ are the semigroups of functions of complexity at most n, i.e. the functions with complexity lower than n are all in separate congruence classes and the functions with complexity greater or equal to n are in the congruence class of 0. The composition of two functions $\phi, \psi \in \mathcal{S}_F/\mathcal{S}_F^n$ with

$comp(\phi), comp(\psi) < n$ gives $\phi \circ \psi$ if $comp(\phi \circ \psi) < n$ and zero otherwise. Thus, the quotient $\mathcal{S}_F/\mathscr{R}_n$ naturally injects into $\mathcal{S}_F/\mathscr{R}_{n+1}$ for $2 \leq n \leq N-1$ as a set

$$\iota_n : \mathcal{S}_F/\mathscr{R}_n \hookrightarrow \mathcal{S}_F/\mathscr{R}_{n+1}.$$

However, this is not a semigroup homomorphism. Furthermore, the congruences $\mathscr{R}_n$ are totally ordered as

$$\mathscr{R}_N < \mathscr{R}_{N-1} < ... < \mathscr{R}_1$$

and give rise to projections

$$\pi_n : \mathcal{S}_F/\mathscr{R}_{n+1} \twoheadrightarrow \mathcal{S}_F/\mathscr{R}_n,$$

where the π_n are semigroup homomorphisms.

The biological interpretation of the quotients $\mathcal{S}_F/\mathscr{R}_n$ now follows immediately: They capture the local structure of the CRS of "size at most n", i.e. within the quotient $\mathcal{S}_F/\mathscr{R}_n$ it is only possible to see those functions that contain reaction pathways of length smaller than n. It is possible to compose the functions as usual, but as soon as the compositions gain a complexity larger than n, the functions vanish, i.e. one is restricted to interactions within "local patches" of limited size. Returning to the idea of relating congruences to coarse-graining schemes, the $\mathscr{R}_n$ describe a rather unusual coarse-graining of the system: Lumping together functions of large complexity can be thought of lumping together "the environment" and retaining the local structure. However, the coarse-graining via the $\mathscr{R}_n$ does not fix a given subnetwork and then integrates out all of its environment, but preserves all the local patches. It is well possible to combine functions in $\mathcal{S}_F/\mathscr{R}_n$ that seemingly live on different patches.

The injections $\iota_n : \mathcal{S}_F/\mathscr{R}_n \hookrightarrow \mathcal{S}_F/\mathscr{R}_{n+1}$ are inclusions of patches of size n into patches of size $n+1$ and the projections $\pi_n : \mathcal{S}_F/\mathscr{R}_{n+1} \twoheadrightarrow \mathcal{S}_F/\mathscr{R}_n$ lose information about functions with complexity $n+1$ and thus correspond to a reduction to smaller patches. This interpretation as a coarse-graining of the environment is illustrated in figure 5.19.

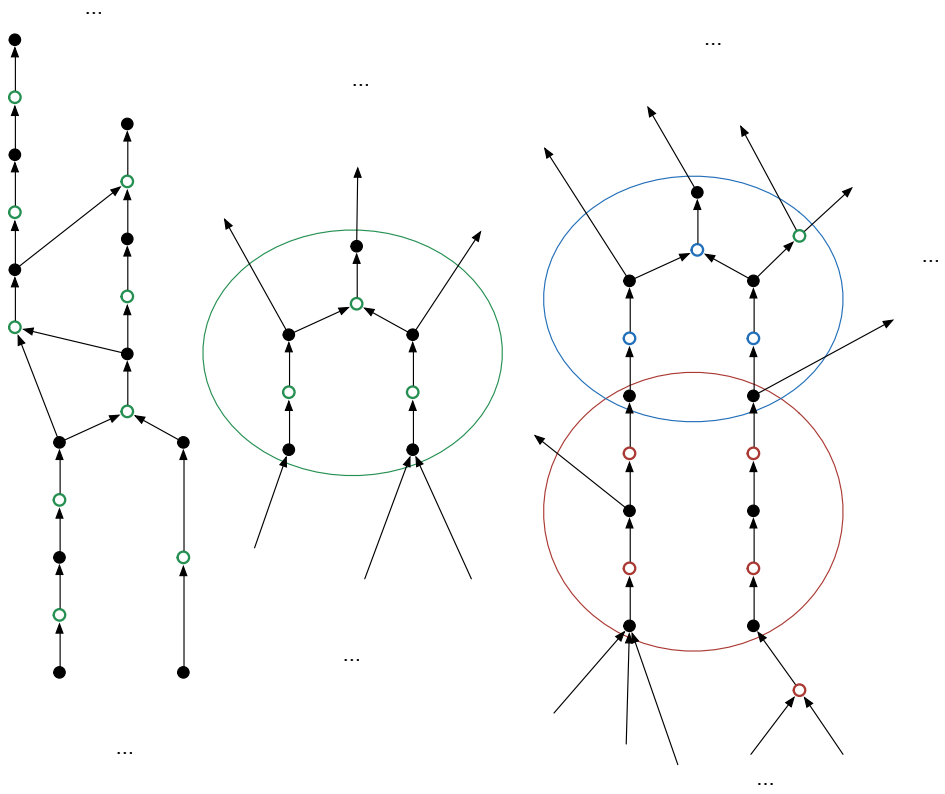


Figure 5.19: Illustration of coarse-graining of the environment via the congruence $\mathscr{R}_3$. The figure shows three functions ϕ_{green}, ϕ_{blue}, ϕ_{red} colored in green, blue and red via the representation of elements in $\mathcal{S}_F$ as pathways in the CRS (remark 5.2.16). The circles indicate the local patches of complexity at most 2. Each of the functions has a local structure of complexity 2 lying in the respective circles. The functions ϕ_{green}, ϕ_{blue}, ϕ_{red} are nonzero in $\mathcal{S}_F/\mathscr{R}_3$. The composition $\phi_{\text{green}} \circ \phi_{\text{blue}}$ gives the function in the blue patch. It has complexity ≤ 2 as well. The composition $\phi_{\text{blue}} \circ \phi_{\text{red}}$ has complexity 4 and equals zero in $\mathcal{S}_F/\mathscr{R}_3$.

Interplay of $\mathscr{B}$ and $\mathscr{R}_n$

After the presentation of the congruence $\mathscr{B}$ on the semigroup of constant functions $\mathcal{S}_c$ of a RAF network, it has been shown that $\mathscr{B}$ cannot be extended to a congruence on $\mathcal{S}_F$. By construction, the congruences $\mathscr{R}_n$ considered in the previous paragraphs contain all constant functions in the congruence class of 0, i.e. have the coarsest possible resolution on the elements of $\mathcal{S}_c$ by lumping them all together into the zero element. The congruences $\mathscr{R}_n$ project to congruences on $\mathcal{S}_F/\mathcal{S}_c$ and thus the congruences $\mathscr{B}$ and $\mathscr{R}_n$ complement each other: While $\mathscr{B}$ contains the global information on all pathways within the network, the $\mathscr{R}_n$ allow to study the local interactions of functions and to disregard functions of too high complexity.

5.7 Discussion

The constructed semigroup models were motivated by the ideas proposed by Oparin and formalized by Kauffman. The original work by Kauffman [190] and subsequent work

within the CRS formalism [217, 198, 218] is aimed primarily at the evaluation of probabilities for the occurrence of RAF subnetworks within a CRS of given size. From a mathematical point of view, this is the task of constructing a map from the set of all CRS to the set $\{0, 1\}$ that takes the value 1 if there is a RAF subnetwork within the respective CRS and 0 otherwise. Such a map is clearly not invertible, i.e. it loses information on the structure of the particular CRS. The models presented here are different in spirit: They retain the full topology of the CRS and do not reduce the information content, i.e. the map assigning the semigroup model $\mathcal{S}_F$ to a CRS is invertible in general.

The models constructed by Rhodes in [212] are very to similar to the semigroup models $\mathcal{S}_F$ proposed here. Rhodes modeled the citric acid cycle with the state space formed by subsets of metabolites involved in the cycle. The semigroup was defined by the actions of all enzymes involved as catalysts in the cycle and all possible compositions thereof. In the language used in this work, he considered a CRS (X, R, C) with state space $X = M \coprod E$ consisting of all metabolites M and enzymes E involved in the cycle such that each reaction $r \in R$ has $dom(r), ran(r) \subset M$ and is catalyzed by some element $e \in E$. The semigroup model $\mathcal{S}_{\text{Rhodes}}$ is generated by all functions $\{\phi_e\}_{e\in E}$ under the operation of composition $\circ$. Rhodes then analyzed the complexity of $\mathcal{S}_{\text{Rhodes}}$ using the Krohn-Rhodes decomposition theorem. Therefore, from a mathematical point of view, the semigroup models $\mathcal{S}_F$ are an extension the semigroups $\mathcal{S}_{\text{Rhodes}}$. Because Rhodes did not allow the enzymes to participate in reactions within the network, the models $\mathcal{S}_{\text{Rhodes}}$ are not applicable to self-referential networks. In particular, $\mathcal{S}_{\text{Rhodes}}$ would always have the empty set as the fixed point in the discrete dynamics. Moreover, Rhodes did not use the operation of addition and thus could not consider joint functions of elements. Without this operation, subnetworks Y of CRS with parallel reaction pathways do not allow to naturally define a function on the network (they do not allow the construction of the maximal element Φ_Y).

In connection to the work of Rhodes, it is useful to note that there is a deep theory on the structure of finite semirings [219]. The semigroups $\mathcal{S}_F$ carry two operations $\circ$ and $+$ that satisfy right-distributivity and an inequality replacing left-distributivity (lemma 5.2.11) making them more general than semirings, which require strict left-distributivity. It would be interesting to study how much of the theory for semirings can be transferred to $\mathcal{S}_F$.

The formalization of the notion of function of elements and subnetworks of a CRS was a primary goal of this work and it has successfully been achieved. It allowed to define a natural dynamics on the state space $\mathfrak{X}_F$ and yielded a simple identification of RAF subnetworks via theorem 5.5.5. Moreover, using the congruence $\mathscr{B}$, it was possible to identify the structure of reaction pathways with the CRS. The notion of complexity of functions in $\mathcal{S}_F$ led to the congruences $\mathscr{R}_n$ and to a new kind of coarse-graining procedure. The corresponding quotient semigroups only see local structures of the CRS and therefore can be though as a coarse-graining applied to the environment. However, this is not a coarse-graining in the classical sense where the fine structure of the environment would be completely deleted leaving only a description of some local patch. In contrast, the coarse-graining by the $\mathscr{R}_n$ retains the information on all local patches.

In remark 5.5.6, a transformation of CRS into classical chemical reaction networks

(CRN) was sketched. This link was used to derive useful restrictions on the structure of physically possible CRS. The reverse transformation of reaction networks into CRS would allow to apply the tools developed here to CRN and in particular to have a new way of coarse-graining procedures based on function. This transformation is not as straightforward as the one in remark 5.5.6. It is currently being addressed by the author.

A main field for applications of semigroup theory are automata theory in theoretical computer science and the theory of formal languages. Automata theory deals with questions of computability and computational complexity. The framework developed here therefore suggests to investigate the computational capabilities of catalytic reaction systems as a future direction of research. The general possibility to consider networks as computational devices was suggested by Mikhailov [220]. Within the theory of formal languages, the lowest class of grammars (regular grammars) according to the Chomsky hierarchy is the class of grammars recognizable by finite-state automata. Such automata can in turn be described by finite semigroups and vice versa. This suggests to study the inverse problem (which finite semigroups can be realized as semigroup models of CRS). Yet, one thing is already clear: Finite CRS have finite semigroup models and are therefore always in the lowest complexity class of formal grammars. Therefore, the more interesting questions in this direction arise for the semigroups of infinite reaction networks and their classification in the Chomsky hierarchy. Such networks should be realized as direct limits of finite networks and the respective semigroups would then be the direct limit of the corresponding finite semigroups. One could also work directly with infinite networks, introducing the semigroups models analogously to the finite case, but the arguments based on finiteness of $\mathcal{S}_F$, $\mathfrak{X}_F$ and X used in many proofs then require modification. An extension of the state space from $\{0,1\}^{X_F}$ to $\mathbb{R}_{\geq 0}^{X_F}$ taking into account the concentrations of the respective species would also lead to infinite semigroups.

Appendix A

Forces and Fluxes in Phenomenological Thermodynamics

This appendix sketches the determination of the entropy production via forces and fluxes in classical nonequilibrium thermodynamics [221] and the connection to stochastic thermodynamics [128].

The internal energy U of a chemical system at equilibrium is given by the Euler equation

$$U = TS - pV + \sum_i \mu_i N_i, \tag{A.0.1}$$

with internal energy U, entropy S, temperature T, pressure p, volume V and chemical potentials μ_i at equilibrium such that the sum runs over all chemical species in the reaction mixture and N_i is the number of molecules of the respective chemical. The differential $\mathrm{d}U$ is given by the Gibbs relation

$$\mathrm{d}U = T\mathrm{d}S - p\mathrm{d}V + \sum_i \mu_i \mathrm{d}N_i. \tag{A.0.2}$$

Here, $\sum_i \mu_i \mathrm{d}N_i$ is the chemical work and $p\mathrm{d}V$ the pressure-volume work except performed by the system.

Assuming that no work is performed by the system, equation A.0.2 simplifies to $\mathrm{d}U = T\mathrm{d}S$ (when work is performed, the equations involve more terms, but the idea of the following derivation remains unaltered). Assuming quasi-stationarity in sufficiently small volume elements V_0, this equation can be rewritten using the energy and entropy densities $u = U/V_0$ and $s = S/V_0$ as

$$\frac{\partial s}{\partial t} = \frac{1}{T}\frac{\partial u}{\partial t}. \tag{A.0.3}$$

Note that the densities u and s have a dependence on spatial coordinates. The entropy production σ is the source term in

$$\frac{\partial s}{\partial t} = \sigma - \mathbf{J_s}, \tag{A.0.4}$$

where $\mathbf{J_s}$ is the entropy flux. The conservation law for internal energy is

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{J_u} = 0, \tag{A.0.5}$$

where $\mathbf{J_u}$ is the flux of internal energy. Using the assumption that no work is performed by the system, this is the heat flux $\mathbf{J_u} = \mathbf{J_q}$ by the first law of thermodynamics. Substituting the conservation law into equation A.0.2 and using the chain rule for $\nabla \cdot (\mathbf{J_q}/T)$ gives

$$\frac{\partial s}{\partial t} = \mathbf{J_q} \cdot \nabla \frac{1}{T} - \nabla \cdot \frac{\mathbf{J_q}}{T}. \tag{A.0.6}$$

From this equation, the entropy flux can be identified as $\mathbf{J_s} = \nabla \cdot \mathbf{J_q}/T$ and comparison with equation A.0.4 gives the entropy production as the the source term

$$\sigma = \mathbf{J_q} \cdot \nabla \frac{1}{T}. \tag{A.0.7}$$

This yields

$$\sigma = -\frac{\mathbf{J_q}}{T^2} \nabla T. \tag{A.0.8}$$

In this system, thermal equilibrium is achieved through heat the transport quantified by $\mathbf{J_q}$. The flux $\mathbf{J_q}$ is conjugate to the force $-1/T^2 \nabla T$ and their product describes the effect of heat transport on the entropy production. Moreover, Fourier's law gives a relationship between the two quantities

$$\mathbf{J_q} = -K \nabla T, \tag{A.0.9}$$

where K is the thermal conductivity. Usually equation A.0.8 includes other pairs of forces and fluxes, originating for example fluxes of chemicals (with force $-\nabla \mu_i$) or of charge (with force $-\nabla \phi$, where ϕ is the electric potential). In these cases the relation between fluxes and forces are given by Fick's and Ohm's laws, respectively. In general, the entropy production can be expressed as a sum of products of conjugate pairs of thermodynamic forces and fluxes

$$\sigma = \sum_{\alpha \in \mathcal{A}} F_\alpha J_\alpha. \tag{A.0.10}$$

The theory of irreversible thermodynamics describes the relationship between fluxes $\mathbf{J}_\alpha$ and forces $\mathbf{F}_\beta$ in the linear regime, i.e. in the regime of small deviations from equilibrium which allows to write

$$\mathbf{J}_\alpha = \sum_\beta L_{\alpha\beta} \mathbf{F}_\beta, \tag{A.0.11}$$

where $L_{\alpha\beta}$ are coupling coefficients [221]. Onsager has derived symmetry relations for the coefficients $L_{\alpha\beta}$ from the microscopic reversibility of underlying processes [222, 223].

Schnakenberg has shown in [128] that the forces $F_\Gamma = \sum_\Gamma F_{x,x'}$ on closed cycles of a Markov network as determined by the equation 1.3.17 in the main text are the *macroscopic* forces generated by a coupling of the system to macroscopic resevoirs. Therefore, the entropy production in a steady-state of the microscopic system considered the main text is actually the entropy production in the macroscopic reservoirs caused by the maintanance of constant potentials. It is given by equation A.0.10. The conjugate fluxes can be determined microscopically in this case by the following procedure.

Let $\{\Gamma_i\}_{i\in I}$ be a basis of cycles for the network, i.e. any cycle on the network can be obtained as a linear combination of the Γ_i with integer coefficients whereby a negative sign indicates reversal of direction and edges with opposite directions cancel when two cycles are added to each other. Such a basis can be obtained as follows: Fix some maximal spanning tree of the network. Each edge from x' to x that belongs to the network, but not to the spanning tree, generates a cycle on the network (because of the maximality of the spanning tree). Denote this cycle by Γ_i and the corresponding probability flux by $J_i := J_{x,x'}$. The choice of direction of the flux J_i defines the direction of the cycle Γ_i. The basis $\{\Gamma_i\}_{i\in I}$ is consists of all the cycles obtained from edges of the network not present in the maximal spanning tree indexed by I; it comes equipped with the set of corresponding probability fluxes $\{J_i\}_{i\in I}$ just defined.

Schnakenberg verified the formula for the total entropy production σ of the network

$$\sigma = \sum_{i\in I} F_{\Gamma_i} J_i. \tag{A.0.12}$$

This expression is formally identical to equation A.0.10, whereby the force F_{Γ_i} are determined by macroscopic reservoirs and the fluxes J_i by the microscopic details of the network. It is a generalization of equation A.0.10 in the sense that it allows to assign an entropy production to microscopic systems with strong fluctuations. When considering a sufficiently large number of copies of the microscopic system as a grand canonical ensemble, the probability fluxes become material fluxes and equation A.0.12 recovers equation A.0.10.

The expression for the entropy production in equation A.0.12 can also be written as [128]

$$\sigma = \frac{1}{2}\sum_{x,x'} J_{x,x'} \ln \frac{w_{x,x'}p(x';t)}{w_{x',x}p(x;t)}. \tag{A.0.13}$$

This is equation 1.3.7 from the main text. It shows that the formula A.0.12 is independent of the choice of cycle basis and motivates the definition of entropy production $\sigma_{x,x'} = J_{x,x'} \ln(w_{x,x'}p(x';t)/w_{x',x}p(x;t))$ for each link on the network. This definition is justified, because the sum of the entropy production of all links recovers the macroscopic entropy production A.0.10 at a steady-state through equation A.0.12. However, the author knows of no physically meaningful way to establish a connection between any individual $\sigma_{x,x'}$ and classical thermodynamical quantities.

Appendix B

Results of Numerical Simulations

B.1 Numerical Results under Experimental Substrate Concentrations

The results of stochastic simulations are graphically presented in the main text. In this section, the numerical values for the respective figures are given. Data for the turnover time distributions (figures 2.7 and 2.11) is given in table B.4, for the stationary probability distributions (figures 2.8 and 2.10) in tables B.1, B.2, B.3 and B.4.

$p(a,b)$	empty	IGP	indole+G3P	G3P
empty	$6.60{\cdot}10^{-2}$	$7.69{\cdot}10^{-2}$	0	$4.41{\cdot}10^{-3}$
Q_1	$2.21{\cdot}10^{-2}$	$1.34{\cdot}10^{-2}$	0	$9.66{\cdot}10^{-4}$
A-A	0	$9.68{\cdot}10^{-3}$	$6.36{\cdot}10^{-3}$	$1.23{\cdot}10^{-2}$
A-A(indole)	0	0	0	$2.63{\cdot}10^{-2}$
Q_3	0	$3.73{\cdot}10^{-2}$	$8.44{\cdot}10^{-2}$	$4.77{\cdot}10^{-1}$
Aex_2	$6.69{\cdot}10^{-2}$	$6.93{\cdot}10^{-2}$	0	$2.68{\cdot}10^{-2}$

Table B.1: Joint probabilities $p(a,b)$ to find the enzyme in the state (a,b).

a	$p(a)$	b	$p(b)$
empty	0.155	empty	0.147
IGP	0.207	Q_1	0.036
indole + G3P	0.091	A-A	0.028
G3P	0.548	A-A(indole)	0.026
		Q_3	0.599
		Aex_2	0.163

Table B.2: Marginal probabilities $p(a)$ and $p(b)$.

$p(a,b)$	empty	IGP	indole+G3P	G3P
empty	$9.24{\cdot}10^{-2}$	$1.11{\cdot}10^{-1}$	0	$2.53{\cdot}10^{-3}$
Q_1	$7.57{\cdot}10^{-2}$	$8.52{\cdot}10^{-2}$	0	$5.98{\cdot}10^{-4}$
A-A	0	$1.59{\cdot}10^{-1}$	$3.80{\cdot}10^{-3}$	$9.50{\cdot}10^{-3}$
A-A(indole)	0	0	0	$1.51{\cdot}10^{-2}$
Q_3	0	$2.23{\cdot}10^{-2}$	$4.90{\cdot}10^{-2}$	$2.79{\cdot}10^{-1}$
Aex_2	$3.89{\cdot}10^{-2}$	$3.99{\cdot}10^{-2}$	0	$1.59{\cdot}10^{-2}$

Table B.3: Simulation setup without activations: Joint probabilities $p(a,b)$ to find the enzyme in the state (a,b).

$p(a,b)$	empty	IGP	indole+G3P	G3P
empty	$1.88{\cdot}10^{-2}$	$2.19{\cdot}10^{-2}$	0	$1.27{\cdot}10^{-3}$
Q_1	$6.50{\cdot}10^{-3}$	$3.92{\cdot}10^{-3}$	0	$2.99{\cdot}10^{-4}$
A-A	0	$2.88{\cdot}10^{-3}$	$1.89{\cdot}10^{-3}$	$4.83{\cdot}10^{-2}$
A-A(indole)	0	0	0	$7.53{\cdot}10^{-3}$
Q_3	0	$1.14{\cdot}10^{-2}$	$6.89{\cdot}10^{-1}$	$1.39{\cdot}10^{-1}$
Aex_2	$1.95{\cdot}10^{-2}$	$2.00{\cdot}10^{-2}$	0	$7.88{\cdot}10^{-3}$

Table B.4: Simulation setup with permanent activations: Joint probabilities $p(a,b)$ to find the enzyme in the state (a,b).

	μ	σ	Q_{25}	Q_{75}
Native enzyme	0.154 s	0.146 s	0.077 s	0.183 s
Permanent activations	0.520 s	1.879 s	0.078 s	0.196 s
Absent activations	0.264 s	0.176 s	0.153 s	0.325 s

Table B.5: Statistical data for simulations with different setups of allosteric activations. μ: Mean turnover time, σ: standard deviation, Q_{25} and Q_{75}: quantiles.

B.2 Numerical Results under Physiological Substrate Concentrations

$\bar{p}(a.b)$	empty	IGP	indole+G3P	G3P
empty	$5.88{\cdot}10^{-1}$	$1.02{\cdot}10^{-1}$	0	$2.99{\cdot}10^{-2}$
Q_1	$4.21{\cdot}10^{-2}$	$2.72{\cdot}10^{-3}$	0	$2.26{\cdot}10^{-3}$
A-A	0	$2.00{\cdot}10^{-3}$	$1.38{\cdot}10^{-3}$	$3.77{\cdot}10^{-2}$
indole+A-A	0	0	0	$1.51{\cdot}10^{-2}$
Q_3	0	$4.45{\cdot}10^{-3}$	$9.70{\cdot}10^{-3}$	$9.81{\cdot}10^{-2}$
Aex_2	$4.95{\cdot}10^{-2}$	$7.98{\cdot}10^{-3}$	0	$7.72{\cdot}10^{-3}$

Table B.6: Stationary probabilities $\bar{p}(a.b)$ to find the enzyme in the state (a,b) under steady-state physiological conditions.

Bibliography

[1] D. Noble. Modeling the heart–from genes to cells to the whole organ. *Science*, 295(5560):1678–1682, 2002.

[2] E. J. Crampin, M. Halstead, P. Hunter, P. Nielsen, D. Noble, N. Smith, and M. Tawhai. Computational physiology and the physiome project. *Experimental Physiology*, 89(1):1–26, 2004.

[3] S. Brenner. Biological computation. *The Limits of Reductionism in Biology*, 213:106–116, 1998.

[4] D. Noble. *The music of life: biology beyond genes*. Oxford University Press, 2008.

[5] L. Nottale. Scale relativity and fractal space-time: applications to quantum physics, cosmology and chaotic systems. *Chaos, Solitons & Fractals*, 7(6):877–938, 1996.

[6] E. Zamir and B. Geiger. Molecular complexity and dynamics of cell-matrix adhesions. *Journal of Cell Science*, 114(20):3583–3590, 2001.

[7] I. Cvitkovic and M. S. Jurica. Spliceosome database: a tool for tracking components of the spliceosome. *Nucleic Acids Research*, 41(D1):D132–D141, 2012.

[8] M. C. Wahl, C. L. Will, and R. Lührmann. The spliceosome: design principles of a dynamic rnp machine. *Cell*, 136(4):701–718, 2009.

[9] Z. H. Zhou, D. B. McCarthy, C. M. O'Connor, L. J. Reed, and J. K. Stoops. The remarkable structural and functional organization of the eukaryotic pyruvate dehydrogenase complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26):14802–14807, 2001.

[10] C. Z. Constantine, C. M. Starks, C. P. Mill, A. E. Ransome, S. J. Karpowicz, J. A. Francois, R. A. Goodman, and T. J. Kappock. Biochemical and structural studies of n 5-carboxyaminoimidazole ribonucleotide mutase from the acidophilic bacterium acetobacter aceti. *Biochemistry*, 45(27):8193–8208, 2006.

[11] S. X. Li, Y. P. Tong, X. C. Xie, Q. H. Wang, H. N. Zhou, Y. Han, Z. Y. Zhang, W. Gao, S. G. Li, X. C. Zhang, et al. Octameric structure of the human bifunctional enzyme PAICS in purine biosynthesis. *Journal of Molecular Biology*, 366(5):1603–1614, 2007.

[12] S. An, R. Kumar, E. D. Sheets, and S. J. Benkovic. Reversible compartmentalization of de novo purine biosynthetic complexes in living cells. *Science*, 320(5872):103–106, 2008.

[13] M. W. Górna, A. J. Carpousis, and B. F. Luisi. From conformational chaos to robust regulation: the structure and function of the multi-enzyme rna degradosome. *Quarterly Reviews of Biophysics*, 45(2):105–145, 2012.

[14] A. M. van Oijen and J. J. Loparo. Single-molecule studies of the replisome. *Annual Review of Biophysics*, 39:429–448, 2010.

[15] R. Nussinov, B. Ma, and C. J. Tsai. A broad view of scaffolding suggests that scaffolding proteins can actively control regulation and signaling of multienzyme complexes through allostery. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1834(5):820–829, 2013.

[16] E. W. Miles, S. Rhee, and D. R. Davies. The molecular basis of substrate channeling. *Journal of Biological Chemistry*, 274(18):12193–12196, 1999.

[17] W. Chuenchor, T. I. Doukov, M. Resto, A. Chang, and B. Gerratana. Regulation of the intersubunit ammonia tunnel in mycobacterium tuberculosis glutamine-dependent NAD+ synthetase. *Biochemical Journal*, 443(2):417–426, 2012.

[18] X. Huang, H. M. Holden, and F. M. Raushel. Channeling of substrates and intermediates in enzyme-catalyzed reactions. *Annual Review of Biochemistry*, 70(1):149–180, 2001.

[19] M. F. Dunn. Allosteric regulation of substrate channeling and catalysis in the tryptophan synthase bienzyme complex. *Archives of Biochemistry and Biophysics*, 519(2):154–66, 2012.

[20] A. Amadasi, M. Bertoldi, R. Contestabile, S. Bettati, B. Cellini, M. L. di Salvo, C. Borri-Voltattorni, F. Bossa, and A. Mozzarelli. Pyridoxal 5'-phosphate enzymes as targets for therapeutic agents. *Current Medicinal Chemistry*, 14(12):1291–1324, 2007.

[21] D. Loutchko, D. Gonze, and A. S. Mikhailov. Single-molecule stochastic analysis of channeling enzyme tryptophan synthase. *Journal of Physical Chemistry B*, 120(9):2179–2186, 2016.

[22] D. Hartich, A. C. Barato, and U. Seifert. Stochastic thermodynamics of bipartite systems: transfer entropy inequalities and a maxwells demon interpretation. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(2):P02016, 2014.

[23] J. M. Horowitz and M. Esposito. Thermodynamics with continuous information flow. *Physical Review X*, 4(3):031015, 2014.

[24] G. Diana and M. Esposito. Mutual entropy production in bipartite systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(4):P04010, 2014.

[25] D. Loutchko, M. Eisbach, and A. S. Mikhailov. Stochastic thermodynamics of a chemical nanomachine: The channeling enzyme tryptophan synthase. *Journal of Chemical Physics*, 146(2):025101, 2017.

[26] W. W. Umbreit, W. A. Wood, and I. C. Gunsalus. The activity of pyridoxal phosphate in tryptophane formation by cell-free enzyme preparations. *The Journal of Biological Chemistry*, 165(2):731, 1946.

[27] I. P. Crawford and C. Yanofsky. On the separation of the tryptophan synthetase of Escherichia Coli Into two protein components. *Proceedings of the National Academy of Sciences of the United States of America*, 44(12):1161–70, 1958.

[28] H. Ngo, R. Harris, N. Kimmich, P. Casino, D. Niks, L. Blumenstein, T. R. Barends, V. Kulik, M. Weyand, I. Schlichting, and M. F. Dunn. *Biochemistry*.

[29] H. Ngo, N. Kimmich, R. Harris, D. Niks, L. Blumenstein, V. Kulik, T. R. Barends, I. Schlichting, and M. F. Dunn. Allosteric regulation of substrate channeling in tryptophan synthase: modulation of the L-serine reaction in stage I of the β-reaction by α-site ligands. *Biochemistry*, 46(26):7740–53, 2007.

[30] T. R. M. Barends, M. F. Dunn, and I. Schlichting. Tryptophan synthase, an allosteric molecular factory. *Current Opinion in Chemical Biology*, 12(5):593–600, 2008.

[31] M. F. Dunn, D. Niks, H. Ngo, T. R. M. Barends, and I. Schlichting. Tryptophan synthase: the workings of a channeling nanomachine. *Trends in Biochemical Sciences*, 33(6):254–64, 2008.

[32] V. Kulik, M. Weyand, R. Seidel, D. Niks, D. Arac, M. F. Dunn, and I. Schlichting. On the Role of αThr183 in the Allosteric Regulation and Catalytic Mechanism of Tryptophan Synthase. *Journal of Molecular Biology*, 324(4):677–690, 2002.

[33] M. Weyand, I. Schlichting, A. Marabotti, and A. Mozzarelli. Crystal structures of a new class of allosteric effectors complexed to tryptophan synthase. *Journal of Biological Chemistry*, 277(12):10647–52, 2002.

[34] M. Weyand, I. Schlichting, P. Herde, A. Marabotti, and A. Mozzarelli. Crystal structure of the βSer178-Pro mutant of tryptophan synthase. A "knock-out" allosteric enzyme. *Journal of Biological Chemistry*, 277(12):10653–60, 2002.

[35] A. Sachpatzidis, C. Dealwis, J. B. Lubetsky, P. Liang, K. S. Anderson, and E. Lolis. Crystallographic studies of phosphonate-based α-reaction transition-state analogues complexed to tryptophan synthase. *Biochemistry*, 38(39):12665–12674, 1999.

[36] A. Marabotti, P. Cozzini, and A. Mozzarelli. Novel allosteric effectors of the tryptophan synthase $\alpha_2\beta_2$ complex identified by computer-assisted molecular modeling. *Biochimica et Biophysica Acta - Protein Structure and Molecular Enzymology*, 1476(2):287–299, 2000.

[37] V. Kulik, E. Hartmann, M. Weyand, M. Frey, A. Gierl, D. Niks, M. F. Dunn, and I. Schlichting. On the structural basis of the catalytic mechanism and the regulation of the α-subunit of tryptophan synthase from Salmonella typhimurium and BX1 from maize, two evolutionarily related enzymes. *Journal of Molecular Biology*, 352(3):608–20, 2005.

[38] T. R .M. Barends, T. Domratcheva, V. Kulik, L. Blumenstein, D. Niks, M. F. Dunn, and I. Schlichting. Structure and mechanistic implications of a tryptophan synthase quinonoid intermediate. *Chembiochem : A European Journal of Chemical Biology*, 9(7):1024–8, 2008.

[39] C. C. Hyde, S. A. Ahmed, E. A. Padlan, E. W. Miles, and D. R. Davies. Three-dimensional structure of the tryptophan synthase $\alpha_2\beta_2$ multienzyme complex from Salmonella typhimurium. *Journal of Biological Chemistry*, 263(33):17857–71, 1988.

[40] E. W. Miles. Structural basis for catalysis by tryptophan synthase. *Advances in Enzymology and Related Areas of Molecular Biology*, 64:93–172, 1991.

[41] P. Pan, E. Woehl, and M. F. Dunn. Protein architecture, dynamics and allostery in tryptophan synthase channeling. *Trends in Biochemical Sciences*, 22(1):22–27, 1997.

[42] P. S. Brzović, K. Ngo, and M. F. Dunn. Allosteric interactions coordinate catalytic activity between successive metabolic enzymes in the tryptophan synthase bienzyme complex. *Biochemistry*, 31(15):3831–9, 1992.

[43] C. A. Leja, E. U. Woehl, and M. F. Dunn. Allosteric linkages between β-site covalent transformations and α-site activation and deactivation in the tryptophan synthase bienzyme complex. *Biochemistry*, 34(19):6552–61, 1995.

[44] E. Woehl and M. F. Dunn. Mechanisms of monovalent cation action in enzyme catalysis: the first stage of the tryptophan synthase β-reaction. *Biochemistry*, 38(22):7118–30, 1999.

[45] E. Woehl and M. F. Dunn. Mechanisms of monovalent cation action in enzyme catalysis: the tryptophan synthase α-, β-, and α-β-reactions. *Biochemistry*, 38(22):7131–41, 1999.

[46] Y. X. Fan, P. McPhie, and E. W. Miles. Guanidine hydrochloride exerts dual effects on the tryptophan synthase $\alpha_2\beta_2$ complex as a cation activator and as a modulator of the active site conformation. *Biochemistry*, 38(24):7881–90, 1999.

[47] L. Blumenstein, T. Domratcheva, D. Niks, H. Ngo, R. Seidel, M. F. Dunn, and I. Schlichting. βQ114N and βT110V mutations reveal a critically important role of the substrate α-carboxylate site in the reaction specificity of tryptophan synthase. *Biochemistry*, 46(49):14100–16, 2007.

[48] S. Rhee. Cryo-crystallography of a true substrate, indole-3-glycerol phosphate, bound to a mutant (αD60N) tryptophan synthase $\alpha_2\beta_2$ complex reveals the correct orientation of active site αGlu49. *Journal of Biological Chemistry*, 273(15):8553–8555, 1998.

[49] M. Weyand and I. Schlichting. Crystal structure of wild-type tryptophan synthase complexed with the natural substrate indole-3-glycerol phosphate. *Biochemistry*, 38(50):16469–80, 1999.

[50] T. R. Schneider, E. Gerhardt, M. Lee, P. H. Liang, K. S. Anderson, and I. Schlichting. Loop closure and intersubunit communication in tryptophan synthase. *Biochemistry*, 37(16):5394–406, 1998.

[51] D. Ferrari, L. H. Yang, E. W. Miles, and M. F. Dunn. βD305A mutant of tryptophan synthase shows strongly perturbed allosteric regulation and substrate specificity. *Biochemistry*, 40(25):7421–32, 2001.

[52] D. Ferrari, D. Niks, L. Yang, E. W. Miles, and M. F. Dunn. Allosteric communication in the tryptophan synthase bienzyme complex: roles of the β-subunit aspartate 305-arginine 141 salt bridge. *Biochemistry*, 42(25):7807–18, 2003.

[53] A. Marabotti, D. De Biase, A. Tramonti, S. Bettati, and A. Mozzarelli. Allosteric communication of tryptophan synthase. Functional and regulatory properties of the βS178P mutant. *Journal of Biological Chemistry*, 276(21):17747–53, 2001.

[54] S. Raboni, S. Bettati, and A. Mozzarelli. Identification of the geometric requirements for allosteric communication between the α- and β-subunits of tryptophan synthase. *Journal of Biological Chemistry*, 280(14):13450–6, 2005.

[55] A. Peracchi, A. Mozzarelli, and G. L. Rossi. Monovalent cations affect dynamic and functional properties of the tryptophan synthase $\alpha_2\beta_2$ complex. *Biochemistry*, 34(29):9459–9465, 1995.

[56] S. Rhee, K. D. Parris, S. A. Ahmed, E. W. Miles, and D. R. Davies. Exchange of K+ or Cs+ for Na+ induces local and long-range changes in the three-dimensional structure of the tryptophan synthase $\alpha_2\beta_2$ complex. *Biochemistry*, 35(13):4211–21, 1996.

[57] A. T. Dierkers, I. Niks, D. Schlichting, and M. F. Dunn. Tryptophan synthase: structure and function of the monovalent cation site. *Biochemistry*, 48(46):10997–10100, 2009.

[58] E. U. Woehl and M. F. Dunn. Monovalent metal ions play an essential role in catalysis and intersubunit communication in the tryptophan synthase bienzyme complex. *Biochemistry*, 34(29):9466–9476, 1995.

[59] Y. X. Fan, P. McPhie, and E. W. Miles. Regulation of tryptophan synthase by temperature, monovalent cations, and an allosteric ligand. Evidence from Arrhenius plots, absorption spectra, and primary kinetic isotope effects. *Biochemistry*, 39(16):4692–4703, 2000.

[60] P. S. Brzović, C. Craig Hyde, Edith W. Miles, and Michael F. Dunn. Characterization of the functional role of a flexible loop in the α-subunit of tryptophan synthase from Salmonella typhimurium by rapid-scanning, stopped-flow spectroscopy and site-directed mutagenesis. *Biochemistry*, 32(39):10404–13, 1993.

[61] R. S. Phillips, P. McPhie, E. W. Miles, S. Marchal, and R. Lange. Quantitative effects of allosteric ligands and mutations on conformational equilibria in Salmonella typhimurium tryptophan synthase. *Archives of Biochemistry and Biophysics*, 470(1):8–19, 2008.

[62] W. F. Drewe and M. F. Dunn. Detection and identification of intermediates in the reaction of L-serine with Escherichia coli tryptophan synthase via rapid-scanning ultraviolet-visible spectroscopy. *Biochemistry*, 24(15):3977–3987, 1985.

[63] R. A. Friesner. Ab initio quantum chemistry: methodology and applications. *Proceedings of the National Academy of Sciences of the United States of America*, 102(19):6648–6653, 2005.

[64] A. Damjanović, H. M. Vaswani, P. Fromme, and G. R. Fleming. Chlorophyll excitations in photosystem I of synechococcus elongatus. *Journal of Physical Chemistry B*, 106(39):10251–10262, 2002.

[65] S. Jurinovich, L. Viani, I. G. Prandi, T. Renger, and B. Mennucci. Towards an ab initio description of the optical spectra of light-harvesting antennae: application to the CP29 complex of photosystem II. *Physical Chemistry Chemical Physics*, 17(22):14405–14416, 2015.

[66] V. I. Novoderezhkin, E. Romero, J. P. Dekker, and R. van Grondelle. Multiple charge-separation pathways in photosystem II: modeling of transient absorption kinetics. *ChemPhysChem*, 12(3):681–688, 2011.

[67] H. Fliegl, K. Fink, W. Klopper, C. E. Anson, A. K. Powell, and R. Clérac. Ab initio study of the magnetic exchange coupling constants of a structural model [CaMn3IIIMnII] of the oxygen evolving center in photosystem II. *Physical Chemistry Chemical Physics*, 11(20):3900–3909, 2009.

[68] S. Sharma, K. Sivalingam, F. Neese, and G. K. L. Chan. Low-energy spectrum of iron–sulfur clusters directly from many-particle quantum mechanics. *Nature Chemistry*, 6(10):927–933, 2014.

[69] M. Radon and K. Pierloot. Binding of CO, NO, and O2 to heme by density functional and multireference ab initio calculations. *Journal of Physical Chemistry A*, 112(46):11824–11832, 2008.

[70] H. Azzouz and D. Borgis. A quantum molecular-dynamics study of proton-transfer reactions along asymmetrical H bonds in solution. *Journal of Chemical Physics*, 98(9):7361–7374, 1993.

[71] S. Hammes-Schiffer. Hydrogen tunneling and protein motion in enzyme reactions. *Accounts of Chemical Research*, 39(2):93–100, 2006.

[72] A. Kuki and P. G. Wolynes. Electron tunneling paths in proteins. *Science*, 236(4809):1647–1652, 1987.

[73] E. Babini, I. Bertini, M. Borsari, F. Capozzi, C. Luchinat, X. Zhang, G. L. C. Moura, I. V. Kurnikov, D. N. Beratan, A. Ponce, et al. Bond-mediated electron tunneling in ruthenium-modified high-potential iron- sulfur protein. *Journal of the American Chemical Society*, 122(18):4532–4533, 2000.

[74] H. B. Gray and J. R. Winkler. Electron tunneling through proteins. *Quarterly Reviews of Biophysics*, 36(3):341–372, 2003.

[75] C. Shih, A. K. Museth, M. Abrahamsson, A. M. Blanco-Rodriguez, A. J. Di Bilio, J. Sudhamsu, B. R. Crane, K. L. Ronayne, M. Towrie, A. Vlček, et al. Tryptophan-accelerated electron flow through proteins. *Science*, 320(5884):1760–1762, 2008.

[76] R. P. Muller and A. Warshel. Ab initio calculations of free energy barriers for chemical reactions in solution. *Journal of Physical Chemistry*, 99(49):17516–17524, 1995.

[77] H. Liu, M. Elstner, E. Kaxiras, T. Frauenheim, J. Hermans, and W. Yang. Quantum mechanics simulation of protein dynamics on long timescale. *Proteins: Structure, Function, and Bioinformatics*, 44(4):484–489, 2001.

[78] P. Carloni, U. Rothlisberger, and M. Parrinello. The role and perspective of ab initio molecular dynamics in the study of biological systems. *Accounts of Chemical Research*, 35(6):455–464, 2002.

[79] E. Rosta, M. Klähn, and A. Warshel. Towards accurate ab initio QM/MM calculations of free-energy profiles of enzymatic reactions. *Journal of Physical Chemistry B*, 110(6):2934–2941, 2006.

[80] H. M. Senn and W. Thiel. QM/MM studies of enzymes. *Current Opinion in Chemical Biology*, 11(2):182–187, 2007.

[81] M. W. van der Kamp and A. J. Mulholland. Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. *Biochemistry*, 52(16):2708–2728, 2013.

[82] Y. Cheng, Y. Zhang, and J. A. McCammon. How does the cAMP-dependent protein kinase catalyze the phosphorylation reaction: an ab initio QM/MM study. *Journal of the American Chemical Society*, 127(5):1553–1562, 2005.

[83] M. W. van der Kamp, F. Perruccio, and A. J. Mulholland. High-level QM/MM modelling predicts an arginine as the acid in the condensation reaction catalysed by citrate synthase. *Chemical Communications*, (16):1874–1876, 2008.

[84] S. Núnez, D. Antoniou, V. L. Schramm, and S. D. Schwartz. Promoting vibrations in human purine nucleoside phosphorylase. a molecular dynamics and hybrid quantum mechanical/molecular mechanical study. *Journal of the American Chemical Society*, 126(48):15720–15729, 2004.

[85] J. C. Schöneboom, H. Lin, N. Reuter, W. Thiel, S. Cohen, F. Ogliaro, and S. Shaik. The elusive oxidant species of cytochrome P450 enzymes: characterization by combined quantum mechanical/molecular mechanical (QM/MM) calculations. *Journal of the American Chemical Society*, 124(27):8142–8151, 2002.

[86] S. C. L. Kamerlin, M. Haranczyk, and A. Warshel. Progress in ab initio QM/MM free-energy simulations of electrostatic energies in proteins: accelerated QM/MM studies of pk_a, redox reactions and solvation free energies. *Journal of Physical Chemistry B*, 113(5):1253–1272, 2008.

[87] M. Karplus and J. Kuriyan. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences of the United States of America*, 102(19):6679–6685, 2005.

[88] R. A. Böckmann and H. Grubmüller. Nanoseconds molecular dynamics simulation of primary mechanical energy transfer steps in F1-ATP synthase. *Nature Structural & Molecular Biology*, 9(3):198–202, 2002.

[89] J. Ma, T. C. Flynn, Q. Cui, A. G. W. Leslie, J. E. Walker, and M. Karplus. A dynamic analysis of the rotation mechanism for conformational change in F1-ATPase. *Structure*, 10(7):921–931, 2002.

[90] W. Yang, Y. Q. Gao, Q. Cui, J. Ma, and M. Karplus. The missing link between thermodynamics and structure in F1-ATPase. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3):874–879, 2003.

[91] G. S. Martin. The road to src. *Oncogene*, 23(48):7910–7917, 2004.

[92] F. Sicheri, I. Moarefi, and J. Kuriyan. Crystal structure of the src family tyrosine kinase hck. *Nature*, 385(6617):602–609, 1997.

[93] X. Wenqing, S. C. Harrison, and M. J. Eck. Three-dimensional structure of the tyrosine kinase c-Src. *Nature*, 385(6617):595–602, 1997.

[94] D. A. Case. Molecular dynamics and NMR spin relaxation in proteins. *Accounts of Chemical Research*, 35(6):325–331, 2002.

[95] T. Hansson and W. F. Oostenbrink, C.and van Gunsteren. Molecular dynamics simulations. *Current Opinion in Structural Biology*, 12(2):190–196, 2002.

[96] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, et al. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

[97] P. L. Freddolino, F. Liu, M. Gruebele, and K. Schulten. Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophysical Journal*, 94(10):75–77, 2008.

[98] T. Veitshans, D. Klimov, and D. Thirumalai. Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Folding and Design*, 2(1):1–22, 1997.

[99] D. L. Coy, M. Wagenbach, and J. Howard. Kinesin takes one 8-nm step for each ATP that it hydrolyzes. *Journal of Biological Chemistry*, 274(6):3667–3671, 1999.

[100] N. Kodera, D. Yamamoto, R. Ishikawa, and T. Ando. Video imaging of walking myosin V by high-speed atomic force microscopy. *Nature*, 468(7320):72–76, 2010.

[101] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski. Coarse-grained protein models and their applications. *Chemical Reviews*, 116(14):7898–7936, 2016.

[102] P. Kar and M. Feig. Recent advances in transferable coarse-grained modeling of proteins. *Advances in Protein Chemistry and Structural Biology*, 96:143–180, 2014.

[103] D. H. de Jong, G. Singh, W. F. D. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schafer, X. Periole, D. P. Tieleman, and S. J. Marrink. Improved parameters for the martini coarse-grained protein force field. *Journal of Chemical Theory and Computation*, 9(1):687–697, 2012.

[104] S. Tanaka and H. A. Scheraga. Medium-and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 9(6):945–950, 1976.

[105] H. Taketomi, Y. Ueda, and N. Gō. Studies on protein folding, unfolding and fluctuations by computer simulation. *Chemical Biology & Drug Design*, 7(6):445–459, 1975.

[106] N. Gō and H. Taketomi. Studies on protein folding, unfolding and fluctuations by computer simulation iv: Hydrophobic interactions. *Chemical Biology & Drug Design*, 13(5):447–461, 1979.

[107] G. G. Maisuradze, P. Senet, C. Czaplewski, A. Liwo, and H. A. Scheraga. Investigation of protein folding by coarse-grained molecular dynamics with the UNRES force field. *Journal of Physical Chemistry A*, 114(13):4471–4485, 2010.

[108] S. Takada. Coarse-grained molecular simulations of large biomolecules. *Current Opinion in Structural Biology*, 22(2):130–137, 2012.

[109] C. Sinner, B. Lutz, S. John, I. Reinartz, A. Verma, and A. Schug. Simulating biomolecular folding and function by native-structure-based/Go-type models. *Israel Journal of Chemistry*, 54(8-9):1165–1175, 2014.

[110] C. Clementi, P. A. Jennings, and J. N. Onuchic. Prediction of folding mechanism for circular-permuted proteins. *Journal of Molecular Biology*, 311(4):879–890, 2001.

[111] C. Clementi, A. E. Garcıa, and J. N. Onuchic. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein L. *Journal of Molecular Biology*, 326(3):933–954, 2003.

[112] L. L. Chavez, J. N. Onuchic, and C. Clementi. Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *Journal of the American Chemical Society*, 126(27):8426–8432, 2004.

[113] M. M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical Review Letters*, 77(9):1905, 1996.

[114] C. Chennubhotla, A. J. Rader, L. W. Yang, and I. Bahar. Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies. *Physical Biology*, 2(4):S173, 2005.

[115] H. Flechsig and A. S Mikhailov. Tracing entire operation cycles of molecular motor hepatitis c virus helicase in structurally resolved dynamical simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 107(49):20875–20880, 2010.

[116] M. Düttmann, Y. Togashi, T. Yanagida, and A. S. Mikhailov. Myosin-v as a mechanical sensor: an elastic network study. *Biophysical Journal*, 102(3):542–551, 2012.

[117] Y. Wang, A. J. Rader, I. Bahar, and R. L. Jernigan. Global ribosome motions revealed with elastic network model. *Journal of Structural Biology*, 147(3):302–314, 2004.

[118] C. Bustamante, D. Keller, and G. Oster. The physics of molecular motors. *Accounts of Chemical Research*, 34(6):412–420, 2001.

[119] S. Liepelt and R. Lipowsky. Kinesins network of chemomechanical motor cycles. *Physical Review Letters*, 98(25):258102, 2007.

[120] E. Zimmermann and U. Seifert. Efficiencies of a molecular motor: a generic hybrid model applied to the F1-ATPase. *New Journal of Physics*, 14(10):103023, 2012.

[121] P. Gaspard and E. Gerritsma. The stochastic chemomechanics of the F1-ATPase molecular motor. *Journal of Theoretical Biology*, 247(4):672–686, 2007.

[122] E. Gerritsma and P. Gaspard. Chemomechanical coupling and stochastic thermodynamics of the F1-ATPase molecular motor with an applied external torque. *Biophysical Reviews and Letters*, 5(04):163–208, 2010.

[123] C. Maes and M. H. van Wieren. A markov model for kinesin. *Journal of Statistical Physics*, 112(1):329–355, 2003.

[124] V. Bierbaum and R. Lipowsky. Chemomechanical coupling and motor cycles of myosin V. *Biophysical Journal*, 100(7):1747–1755, 2011.

[125] D. Tsygankov, A. W. R. Serohijos, N. V. Dokholyan, and T. C. Elston. A physical model reveals the mechanochemistry responsible for dynein's processive motion. *Biophysical Journal*, 101(1):144–150, 2011.

[126] M. A. B. Baker and R. M. Berry. An introduction to the physics of the bacterial flagellar motor: a nanoscale rotary electric motor. *Contemporary Physics*, 50(6):617–632, 2009.

[127] D. Chowdhury. Stochastic mechano-chemical kinetics of molecular motors: a multidisciplinary enterprise from a physicists perspective. *Physics Reports*, 529(1):1–197, 2013.

[128] J. Schnakenberg. Network theory of microscopic and macroscopic behavior of master equation systems. *Reviews of Modern Physics*, 48(4):571, 1976.

[129] K. Sekimoto. Kinetic characterization of heat bath and the energetics of thermal ratchet models. *Journal of the Physical Society of Japan*, 66(5):1234–1237, 1997.

[130] D. J. Evans and D. J. Searles. Equilibrium microstates which generate second law violating steady states. *Physical Review E*, 50(2):1645, 1994.

[131] G. Gallavotti and E. G. D. Cohen. Dynamical ensembles in nonequilibrium statistical mechanics. *Physical Review Letters*, 74(14):2694, 1995.

[132] J. L. Lebowitz and H. Spohn. A gallavotti–cohen-type symmetry in the large deviation functional for stochastic dynamics. *Journal of Statistical Physics*, 95(1):333–365, 1999.

[133] C. Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690–2693, 1997.

[134] G. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60(3):2721–2726, 1999.

[135] G. Hummer and A. Szabo. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 98(7):3658–61, 2001.

[136] T. Hatano and S. I. Sasa. Steady-state thermodynamics of langevin systems. *Physical Review Letters*, 86(16):3463–3466, 2001.

[137] U. Seifert. Entropy Production along a Stochastic Trajectory and an Integral Fluctuation Theorem. *Physical Review Letters*, 95(4):040602, 2005.

[138] C. Van den Broeck. The many faces of the second law. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(10):P10009, 2010.

[139] S. Ito and T. Sagawa. Information thermodynamics on causal networks. *Physical Review Letters*, 111(18):180603, 2013.

[140] J. Hoppenau and A. Engel. On the energetics of information exchange. *Europhysics Letters*, 105(5):6, 2014.

[141] U. Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75(12):126001, 2012.

[142] K. Sekimoto. *Stochastic energetics*. Springer, 2010.

[143] T. Schmiedl and U. Seifert. Stochastic thermodynamics of chemical reaction networks. *Journal of Chemical Physics*, 126(4), 2007.

[144] T. Schmiedl, T. Speck, and U. Seifert. Entropy production for mechanically or chemically driven biomolecules. *Journal of Statistical Physics*, 128(1-2):77–93, 2007.

[145] M. Polettini. Nonequilibrium thermodynamics as a gauge theory. *Europhysics Letters*, 97(3):30003, 2012.

[146] M. Polettini and M. Esposito. Irreversible thermodynamics of open chemical networks. I. Emergent cycles and broken conservation laws. *Journal of Chemical Physics*, 141(2), 2014.

[147] B. Altaner and J. Vollmer. Fluctuation-preserving coarse graining for biochemical systems. *Physical Review Letters*, 108(22):1–5, 2012.

[148] K. Kawaguchi, S. I. Sasa, and T. Sagawa. Nonequilibrium dissipation-free transport in F1-ATPase and the thermodynamic role of asymmetric allosterism. *Biophysical Journal*, 106(11):2450–2457, 2014.

[149] T. Shibata and S. I. Sasa. Equilibrium chemical engines. 67:2666–2670, 1998.

[150] M. Esposito, R. Kawai, K. Lindenberg, and C. van den Broeck. Efficiency at maximum power of low-Dissipation carnot engines. *Physical Review Letters*, 105(October):1–4, 2010.

[151] U. Seifert. Stochastic thermodynamics of single enzymes and molecular motors. *European Physical Journal E*, 34(3):26, 2011.

[152] C. van den Broeck, N. Kumar, and K. Lindenberg. Efficiency of isothermal molecular machines at maximum power. *Physical Review Letters*, 108(21):1–5, 2012.

[153] F. Jülicher, A. Ajdari, and J. Prost. Modeling molecular motors. *Reviews of Modern Physics*, 69(4):1269–1282, 1997.

[154] R. D. Astumian. Thermodynamics and kinetics of molecular motors. *Biophysical Journal*, 98(11):2401–2409, 2010.

[155] A. W. C. Lau, D. Lacoste, and K. Mallick. Nonequilibrium fluctuations and mechanochemical couplings of a molecular motor. *Physical Review Letters*, 99(15):158102, 2007.

[156] D. Andrieux and P. Gaspard. Fluctuation theorems and the nonequilibrium thermodynamics of molecular motors. *Physical Review E*, 74(1):011906, 2006.

[157] M. E. Fisher and A. B. Kolomeisky. Simple mechanochemistry describes the dynamics of kinesin molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 98(14):7748–53, 2001.

[158] J. E. Baker. Free energy transduction in a chemical motor model. *Journal of Theoretical Biology*, 228(4):467–476, 2004.

[159] R. Lipowsky and S. Liepelt. Chemomechanical coupling of molecular motors: thermodynamics, network representations, and balance conditions. *Journal of Statistical Physics*, 130(1):39–67, 2008.

[160] S. Liepelt and R. Lipowsky. Operation modes of the molecular motor kinesin. *Physical Review E*, 79(1):011917, 2009.

[161] J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, and C. Bustamante. Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski's equality. *Science*, 296(5574):1832–1835, 2002.

[162] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco, and C. Bustamante. Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. *Nature*, 437(7056):231–4, 2005.

[163] A. Imparato, S. Luccioli, and A. Torcini. Reconstructing the free-energy landscape of a mechanically unfolded model protein. *Physical Review Letters*, 99(16):168101, 2007.

[164] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa. Thermodynamics of information. *Nature Physics*, 11(2):131–139, 2015.

[165] A. C. Barato and U. Seifert. Unifying three perspectives on information processing in stochastic thermodynamics. *Physical Review Letters*, 112(9):090601, 2014.

[166] D. Mandal and C. Jarzynski. Work and information processing in a solvable model of maxwells demon. *Proceedings of the National Academy of Sciences of the United States of America*, 109(29):11641–11645, 2012.

[167] N. Shiraishi and T. Sagawa. Fluctuation theorem for partially masked nonequilibrium dynamics. *Physical Review E*, 91(1):012130, 2015.

[168] A. N. Lane and K. Kirschner. The catalytic mechanism of tryptophan synthase from Escherichia coli. *European Journal of Biochemistry*, 129(3):571–582, 1983.

[169] K. S. Anderson, E. W. Miles, and K. A. Johnson. Serine modulates substrate channeling in tryptophan synthase. *Journal of Biological Chemistry*, 266(13):8020–8033, 1991.

[170] R. Y. Wang. Rapid scan, stoppedflow kinetics. In *Encyclopedia of Inorganic and Bioinorganic Chemistry*. 2011.

[171] B. A. Barshop, R. F. Wrenn, and C. Frieden. Analysis of numerical methods for computer simulation of kinetic processes: development of KINSIM - a flexible, portable system. *Analytical Biochemistry*, 130(1):134–45, 1983.

[172] A. N. Lane and K. Kirschner. The mechanism of tryptophan binding to tryptophan synthase from Escherichia coli. *European Journal of Biochemistry*, 120(2):379–87, 1981.

[173] A. N. Lane and K. Kirschner. The mechanism of binding of L-serine to tryptophan synthase from Escherichia coli. *European Journal of Biochemistry*, 129(3):561–570, 1983.

[174] C. Yanofsky and M. Rachmeler. The exclusion of free indole as an intermediate in the biosynthesis of tryptophan in Neurospora crassa. *Biochimica et Biophysica Acta*, 28(3):640–1, 1958.

[175] M. F. Dunn, V. Aguilar, P. Brzović, W. F. Drewe, K. F. Houben, C. A. Leja, and M. Roy. The tryptophan synthase bienzyme complex transfers indole between the alpha- and beta-sites via a 25-30 A long tunnel. *Biochemistry*, 29(2):8598–8607, 1990.

[176] K. Kirschner, A. N. Lane, and A. W. Strasser. Reciprocal communication between the lyase and synthase active sites of the tryptophan synthase bienzyme complex. *Biochemistry*, 30(2):472–8, 1991.

[177] A. N. Lane and K. Kirschner. Mechanism of the physiological reaction catalyzed by tryptophan synthase from escherichia coli. *Biochemistry*, 30(2):479–484, 1991.

[178] P. S. Brzović, Y. Sawa, C. C. Hyde, E. W. Miles, and M. F. Dunn. Evidence that mutations in a loop region of the alpha-subunit inhibit the transition from an open to a closed conformation in the tryptophan synthase bienzyme complex. *Journal of Biological Chemistry*, 267(18):13028–38, 1992.

[179] W. F. Drewe and M. F. Dunn. Characterization of the reaction of L-serine and indole with Escherichia coli tryptophan synthase via rapid-scanning ultraviolet-visible spectroscopy. *Biochemistry*, 25(9):2494–501, 1986.

[180] E. J. Faeder and G. G. Hammes. Kinetic studies of tryptophan synthetase. Interaction of substrates with the B subunit. *Biochemistry*, 9(21):4043–9, 1970.

[181] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

[182] S. Raboni, S. Bettati, and A. Mozzarelli. Tryptophan synthase: A mine for enzymologists. *Cellular and Molecular Life Sciences*, 66(14):2391–2403, 2009.

[183] L. Edman, U. Mets, and R. Rigler. Conformational transitions monitored for single molecules in solution. *Proceedings of the National Academy of Sciences of the United States of America*, 93(13):6710–5, 1996.

[184] H. P. Lu. Single-molecule enzymatic dynamics. *Science*, 282(5395):1877–1882, 1998.

[185] T. Ha, A. Y. Ting, J. Liang, W. B. Caldwell, A. A. Deniz, D. S. Chemla, P G Schultz, and S Weiss. Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 96(3):893–8, 1999.

[186] N. Kishore, Y. B. Tewari, D. L. Akers, R. N. Goldberg, and E. W. Miles. A thermodynamic investigation of reactions catalyzed by tryptophan synthase. *Biophysical chemistry*, 73(3):265–280, 1998.

[187] B. D. Bennett, E. H. Kimball, M. Gao, R. Osterhout, S. J. van Dien, and J. D. Rabinowitz. Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. *Nature Chemical Biology*, 5(8):593–599, 2009.

[188] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa. Thermodynamics of information. *Nature Physics*, 11(2):131, 2015.

[189] W. Hordijk and M. Steel. Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *Journal of Theoretical Biology*, 227(4):451–461, 2004.

[190] S. A. Kauffman. Autocatalytic sets of proteins. *Journal of Theoretical Biology*, 119(1):1–24, 1986.

[191] E. F. Keller. *Contemporary debates in philosophy of biology*. John Wiley & Sons, 2009.

[192] J. M. Howie. *Fundamentals of semigroup theory.* Academic Press Inc., 1976.

[193] A. I. Oparin. *The origin of life on the earth.* Oliver & Boyd, Edinburgh & London, 1957.

[194] F. Dyson. *Origins of life.* Cambridge University Press, 1999.

[195] N. R. Pace. Origin of life-facing up to the physical setting. *Cell*, 65(4):531–533, 1991.

[196] D. W. Deamer. Origins of life: How leaky were primitive cells? *Nature*, 454(7200):37–38, 2008.

[197] A. Y. Mulkidjanian, A. Y. Bychkov, D. V. Dibrova, M. Y. Galperin, and E. V. Koonin. Origin of first cells at terrestrial, anoxic geothermal fields. *Proceedings of the National Academy of Sciences of the United States of America*, 109(14):821–830, 2012.

[198] W. Hordijk, M. Steel, and S. Kauffman. The structure of autocatalytic sets: Evolvability, enablement, and emergence. *Acta Biotheoretica*, 60(4):379–392, 2012.

[199] J. von Neumann. The general and logical theory of automata. *Cerebral Mechanisms in Behavior*, 1(41):1–2, 1951.

[200] J. von Neumann and A. W. Burks. *Theory of self-reproducing automata.* University of Illinois Press Urbana, 1966.

[201] J. von Neumann. *The computer and the brain.* Yale Univesity Press, 1958.

[202] N. Rashevsky. Note on the mathematical theory of oxygen consumption at low oxygen pressures. *Protoplasma*, 20(1):125–130, 1933.

[203] N. Rashevsky. Outline of a physico-mathematical theory of excitation and inhibition. *Protoplasma*, 20(1):42–56, 1933.

[204] N. Rashevsky. Further contributions to the theory of cell polarity and self-regulation. *Bulletin of Mathematical Biophysics*, 2(2):65–67, 1940.

[205] N. Rashevsky. The geometrization of biology. *Bulletin of Mathematical Biophysics*, 18(1):31–56, 1956.

[206] N. Rashevsky. Contributions to relational biology. *Bulletin of Mathematical Biology*, 22(1):73–84, 1960.

[207] R. Rosen. A relational theory of biological systems. *Bulletin of Mathematical Biology*, 20(3):245–260, 1958.

[208] R. Rosen. Some realizations of (M, R)-systems and their interpretation. *Bulletin of Mathematical Biology*, 33(3):303–319, 1971.

[209] D. D. Bonchev and O. G. Mekenyan. *Graph theoretical approaches to chemical reactivity.* Springer Science & Business Media, 2012.

[210] F. Noé and C. Clementi. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Current Opinion in Structural Biology*, 43:141–147, 2017.

[211] J. Almeida. *Finite semigroups and universal algebra*, volume 3. World Scientific, 1995.

[212] J. Rhodes. *Applications of automata theory and algebra: via the mathematical theory of complexity to biology, physics, psychology, philosophy, and games*. World Scientific, 2009.

[213] S. Satoh, K. Yama, and M. Tokizawa. Semigroups of order 8. In *Semigroup Forum*, volume 49, pages 7–29. Springer, 1994.

[214] T. L Hill. *Free energy transduction and biochemical cycle kinetics*. Courier Corporation, 2004.

[215] T. Yamura. Indecomposable completely simple semigroups except groups. *Osaka Mathematical Journal*, 8:35–42, 1956.

[216] D. Rees. On semi-groups. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 36, pages 387–400. Cambridge University Press, 1940.

[217] W. Hordijk, S. A. Kauffman, and M. Steel. Required levels of catalysis for emergence of autocatalytic sets in models of chemical reaction systems. *International Journal of Molecular Sciences*, 12(5):3085–3101, 2011.

[218] M. Steel, W. Hordijk, and J. Smith. Minimal autocatalytic networks. *Journal of Theoretical Biology*, 332:96–107, 2013.

[219] J. Rhodes and B. Steinberg. *The q-theory of finite semigroups*. Springer Science & Business Media, 2009.

[220] A. S. Mikhailov. Simple models for complex systems. Workshop: New Frontiers in Nonlinear Sciences, Niseko, 2016.

[221] S. R. De Groot and P. Mazur. *Non-equilibrium thermodynamics*. Dover Books on Physics, 1985.

[222] L. Onsager. Reciprocal relations in irreversible processes. I. *Physical Review*, 37(4):405, 1931.

[223] L. Onsager. Reciprocal relations in irreversible processes. II. *Physical Review*, 38(12):2265, 1931.

Erklärung

Sehr geehrte Damen und Herren,

hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst und keine anderen Hilfsmittel als angegeben verwendet habe. Insbesondere versichere ich, dass ich alle wörtlichen und sinngemen Übernahmen aus anderen Werken als solche kenntlich gemacht habe.

Ich versichere außerdem, dass ich die beigefügte Dissertation nur in diesem und keinem anderen Promotionsverfahren eingereicht habe und, dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind.

Berlin, den 09.02.2018

Dimitri Loutchko